Towards Effective Use of Training Data in Statistical Machine Translation

Philipp Koehn and Barry Haddow
University of Edinburgh
Edinburgh, United Kingdom
{pkoehn,bhaddow}@inf.ed.ac.uk

Abstract

We report on findings of exploiting large data sets for translation modeling, language modeling and tuning for the development of competitive machine translation systems for eight language pairs.

1 Introduction

We report on experiments carried out for the development of competitive systems on the datasets of the 2012 Workshop on Statistical Machine Translation. Our main focus was directed on the effective use of all the available training data during training of translation and language models and tuning.

We use the open source machine translation system Moses (Koehn et al., 2007) and other standard open source tools, hence all our experiments are straightforwardly replicable[1].

Compared to all single system submissions by participants of the workshop we achieved the best BLEU scores for four language pairs (es-en, en-es, cs-en, en-cs), the 2^{nd} best results for two language pairs (fr-en, de-en), as well as a 3^{rd} place (en-de) and a 5^{th} place (en-fr) for the remaining pairs. We improved upon this in the post-evaluation period for some of the language pairs by more systematically applying our methods.

During the development of our system, we saw most gains from using large corpora for translation model training, especially when using subsampling techniques for out-of-domain sets, using large corpora for language model training, and larger tuning sets. We also observed mixed results with alternative tuning methods. We also experimented with hierarchical models and semi-supervised training, but did not achieve any improvements.

[1]Configuration files and instructions are available at http://www.statmt.org/wmt12/uedin/.

LP	Baseline	+UN
fr-en	28.2	28.4 (+.2)
es-en	29.1	28.9 (–.2)
en-fr	28.8	28.7 (–.1)
en-es	31.0	30.9 (–.1)
LP	**Baseline**	**+GigaFrEn**
fr-en	28.7	29.1 (+.4)
en-fr	29.3	30.3 (+1.0)

Table 1: Gains from larger translation models: UN (about 300 million English words), GigaFrEn (about 550 million English words).

We report all results in case-sensitive BLEU (mteval13a) on the newstest2011 test set (Callison-Burch et al., 2011). Please also note that baseline scores vary throughout the paper, since different methods were investigated at different time points.

2 Better Translation Models

2.1 Using Large Training Sets

The WMT evaluation campaign works with the largest training sets in the field. Our French-English systems are trained on a parallel corpus with 1,072 million French and 934 million English words. Training a system on this amount of data takes about two weeks.

The basic data sets for the language pairs are the Europarl and NewsCommentary corpora consist of about 50 million words and 3 million words, respectively. These corpora are quite close to the target domain of news reports, and give quite good results. Table 1 shows the gains from using the much larger UN (about 300 million words) and GigaFrEn corpora (about 550 million words).

From these results, it is not clear if the UN is helpful, but the GigaFrEn corpus gives large gains (+0.4 BLEU and +1.0 BLEU).

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 317–321,
Montréal, Canada, June 7-8, 2012.

LP	Base-line	Model 1				Moore-Lewis			
		Before		After		Before		After	
		10%	50%	10%	50%	10%	50%	10%	50%
fr-en	29.3	28.5(–.8)	29.1(–.2)	28.6(–.7)	28.9(–.4)	29.1(–.2)	**29.6(+.3)**	29.1(–.2)	29.4(+.1)
en-fr	30.1	29.1(–1.0)	30.1(±.0)	29.3(–.8)	29.8(–.3)	29.9(–.2)	**30.2(+.1)**	29.9(–.2)	30.1(±.0)
es-en	29.0	28.9(–.1)	29.0(±.0)	29.0(±.0)	29.0(±.0)	29.0(±.0)	29.1(+.1)	**29.4(+.4)**	29.2(+.2)
en-es	30.9	30.9(±.0)	31.0(+.1)	30.8(–.1)	30.7(–.2)	31.4(+.5)	**31.5(+.6)**	**31.5(+.6)**	31.3(+.4)

Table 2: Subsampling UN and GigaFrEn corpora using Model 1 and Moore-Lewis filtering, before and after word alignment

2.2 Subsampling

We experimented with two different types of subsampling techniques – Model 1, similar to that used by Schwenk et al. (2011), and modified Moore-Lewis (Axelrod et al., 2011) – for the language pairs es-en, en-es, fr-en and en-fr. In each case the idea was to include the NewsCommentary and Europarl corpora in their entirety, and to score the sentences in the remaining corpora (the selection corpus) using one of the two measures, adding either the top 10% or top 50% of the selection corpus to the training data.

For Model 1 filtering, we trained IBM Model 1 on Europarl and NewsCommentary concatenated, in both directions, and scored the sentences in the selection corpus using the length-normalised sum of the IBM Model scores. For the modified Moore-Lewis filtering, we trained two 5-gram language models for source and target, the first on 5M sentences from the news2011 monolingual data, and the second on 5M words from the selection corpus, using the same vocabulary. The modified Moore-Lewis score for a sentence is the sum of the source and target's perplexity difference for the two language models.

For the Spanish experiments, the selection corpus was the UN data, whilst for the French experiments it was the UN data and the GigaFrEn data, concatenated and with duplicates removed.

The results of the subsampling are shown in Table 2, where the BLEU scores are averaged over 2 tuning runs. The conclusion was that modified Moore-Lewis subsampling was effective (and was used in our final submissions), but Model 1 sampling made no difference for the Spanish systems, and was harmful for the French systems.

3 Better Language Models

In previous years, we were not able to make use of the monolingual LDC Gigaword corpora due to lack of sufficiently powerful computing resources. These corpora exist for English (4.3 billion words), Spanish (1.1 billion words), and French (0.8 billion words). With the acquisition of large memory machines[2], we were now able to train language models on this data. Use of these large language models during decoding is aided by more efficient storage and inference (Heafield, 2011).

Still, even with that much RAM it is not possible to train a language model with SRILM (Stolke, 2002) in one pass. Hence, we broke up the training corpus by source (*New York Times*, *Washington Post*, ...) and trained separate language model for each. The largest individual corpus was the English *New York Times* portion which consists of 1.5 billion words and took close to 100GB of RAM. We also trained individual language models for each year of WMT12's monolingual corpus.

We interpolated the language models using the SRILM toolkit. The toolkit has a limit of 10 language models to be merged at once, so we had to interpolate sub-groups of some of the language models (the WMT12 monolingual news models) first. It is not clear if this is harmful, but building separate language model for each source and year and interpolate those many more models did hurt significantly.

Table 3 shows that we gain around half a BLEU point into Spanish and French, as well as German–English, and around one and a half BLEU points for the other language pairs into English.

[2]Dell Poweredge R710, equipped with two 6-core Intel Xeon X5660 CPUs running at 2.8GHz, with each core able to run two threads (24 threads total), six 3TB disks and 144GB RAM, and cost £6000.

LP	Baseline	+LDC Giga
de-en	21.9	22.4 (+.5)
cs-en	24.2	25.6 (+1.4)
fr-en	29.1	31.0 (+1.9)
es-en	29.1	30.7 (+1.6)
en-es	31.5	31.8 (+.3)
en-fr	30.3	30.8 (+.5)

Table 3: Using the LDC Gigaword corpora to train larger language models.

LP	Baseline	Big-Tune
de-en	21.4	21.6 (+.2)
fr-en	28.4	28.7 (+.3)
es-en	28.9	29.0 (+.1)
cs-en	23.9	24.1 (+.2)
en-de	15.8	15.9 (+.1)
en-fr	28.7	29.2 (+.5)
en-es	30.9	31.2 (+.2)
en-cs	17.2	17.4 (+.2)

Table 4: Using a larger tuning set (7567 sentences) by combining newstest 2008 to 2010.

4 Better Tuning

4.1 Bigger Tuning Sets

In recent experiments, mainly geared towards using much larger feature sets, we learned that larger tuning sets may give better and more stable results. We tested this hypothesis here as well.

By concatenating the sets from three years (2008-2010), we constructed a tuning set of 7567 sentences per language. Table 4 shows that we gain on average about +0.2 BLEU points.

4.2 Pairwise Ranked Optimization

We recently added an implementation of the pairwise ranked optimization (PRO) tuning method (Hopkins and May, 2011) to Moses as an alternative to Och's (2003) minimum error rate training (MERT). We checked if this method gives us better results. Table 5 shows a mixed picture. PRO gives slightly shorter translations, probably because it optimises sentence rather than corpus BLEU, which has a noticeable effect on the BLEU score. For 2 language pairs we see better results, for 4 worse, and for 1 there is no difference. On other data and language conditions we have observed better and more stable results with PRO.

LP	MERT	PRO	PRO-MERT
de-en	21.7 (1.01)	21.9 (1.00) +.2	21.7 (1.01) ±.0
es-en	29.1 (1.02)	29.1 (1.01) ±.0	29.1 (1.02) ±.0
cs-en	24.2 (1.03)	24.5 (1.00) +.3	24.2 (1.03) ±.0
en-de	16.0 (1.00)	15.7 (0.96) –.3	16.0 (1.00) ±.0
en-fr	29.3 (0.98)	28.9 (0.96) –.4	29.3 (0.98) ±.0
en-es	31.5 (0.98)	31.3 (0.97) –.2	31.4 (0.98) –.1
en-cs	17.4 (0.97)	16.9 (0.92) –.5	17.3 (0.97) –.1

Table 5: Replacing the line search method of MERT with pairwise ranked optimization (PRO).

We tried to use PRO to generate starting points for MERT optimization. Theoretically this will lead to better optimization on the tuning set, since MERT optimization steps on PRO weights will never lead to worse results on the sampled n-best lists. This method (PRO-MERT in the table) applied here, however, did not lead to significantly different results than plain MERT.

5 What did not Work

Not everything we tried worked out. Notably, two promising directions — hierarchical models and semi-supervised learning — did not yield any improvements. It is not clear if we failed or if the methods failed, but we will investigate this further in future work.

5.1 Hierarchical Models

Hierarchical models (Chiang, 2007) have been supported already for a few years by Moses, and they give significantly better performance for Chinese–English over phrase-based models. While we have not yet seen benefits for many other language pairs, the eight language pairs of WMT12 allowed us to compare these two models more extensively, also in view of recent enhancements resulting in better search accuracy.

Since hierarchical models are much larger (roughly 10 times bigger), we trained hierarchical models on downsized training data for most language pairs. For Spanish and French, this excludes UN and GigaFrEn; for Czech some parts of the CzEng corpus were excluded based on their lower language model interpolation weights relative

LP	Phrase	Downsized	Hierarchical
de-en	21.6	same	21.4 (–.2)
fr-en	28.7	27.9	27.6 (–.3)
es-en	29.0	28.9	28.4 (–.5)
cs-en	24.1	22.4	22.0 (–.4)
en-de	15.9	same	15.5 (–.4)
en-fr	29.2	28.8	28.0 (–.8)
en-es	31.2	30.8	30.4 (–.4)
en-cs	17.4	16.2	15.6 (–.6)

Table 6: Hierarchical phrase models vs. baseline phrase-based models.

to their size.

Table 6 shows inferior performance for all language pairs (by about half a BLEU point), although results for German–English are close (–0.2 BLEU).

5.2 Semi-Supervised Learning

Other research groups have reported improvements using semi-supervised learning methods to create synthetic parallel data from monolingual data (Schwenk et al., 2008; Abdul-Rauf and Schwenk, 2009; Bertoldi and Federico, 2009; Lambert et al., 2011). The idea is to translate in-domain monolingual data with a baseline system and filter the result for use as an additional parallel corpus.

Table 7 shows out results when trying to emulate the approach of Lambert et al. (2011). We translate the some of the 2011 monolingual news data (139 million words for French and 100 million words for English) from the target language into the source language with a baseline system trained on Europarl and News Commentary. Adding all the obtained data hurts (except for minimal improvements over a small French-English system). When we filtered out half of the sentences based on translation scores, results were even worse.

Setup	Baseline	+synthetic	+syn-half
fr-en ep+nc	28.0	28.1 (+.1)	28.0 (±.0)
+un	28.7	28.6 (–.1)	28.5 (–.2)
en-fr ep+nc	28.8	28.2 (–.6)	28.1 (–.7)
+un	29.3	28.9 (–.4)	28.9 (–.4)

Table 7: Using semi-supervised methods to add synthetic parallel data to a baseline system trained on Europarl (ep)m News Commentary (nc) and United Nations (un). We added all generated data (synthetic) or filtered out half based on model scores (syn-half).

Acknowledgments

This work was supported by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme).

References

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece. Association for Computational Linguistics.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Lambert, P., Schwenk, H., Servan, C., and Abdul-Rauf, S. (2011). Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, Scotland. Association for Computational Linguistics.

Och, F. J. (2003). Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*.

Schwenk, H., Estève, Y., and Rauf, S. A. (2008). The LIUM Arabic/English statistical machine translation system for IWSLT 2008. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 63–68.

Schwenk, H., Lambert, P., Barrault, L., Servan, C., Abdul-Rauf, S., Afli, H., and Shah, K. (2011). Lium's smt machine translation systems for wmt 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 464–469, Edinburgh, Scotland. Association for Computational Linguistics.

Stolke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*.