Fine-tuning in Portuguese-English Statistical Machine Translation

Wilker F. Aziz, Thiago A. S. Pardo [1], Ivandré Paraboni [2]

[1]NILC/ICMC, University of São Paulo
Av. Trabalhador São-Carlense, 400, São Carlos, Brazil

[2]EACH, University of São Paulo
Av. Arlindo Bettio, 1000, São Paulo, Brazil

wilker.aziz@usp.br, taspardo@icmc.usp.br, ivandre@usp.br

Abstract. *In previous work we have shown results of a first experiment in Statistical Machine Translation (SMT) for Brazilian Portuguese and American English using state-of-the-art phrase-based models. In this paper we compare a number of training and decoding parameter choices for fine-tuning the system as an attempt to obtain optimal results for this language pair.*

1. Introduction

In recent years research on Statistical Machine Translation (SMT) has witnessed major improvements in translation quality with the use of *phrase-based* techniques, i.e., the use of translation models that take into account the alignment of arbitrary sequences of words that may be linguistically-motivated or not (Koehn et al., 2003). To achieve optimal results, systems of this kind can be customised to a particular language pair or domain, which involves a choice of appropriate alignment heuristics, decoding algorithm, language models of suitable order, and a large number of configuration options for all applicable resources.

In previous work we have presented a number of experiments in SMT in which some of these options have been adjusted to Brazilian Portuguese/American English translation. Additional fine-tuning is both possible and desirable, but this has to be balanced against training time constraints. In our simple system, for example, the training task may take from 30 minutes to 15 hours on a standard hardware configuration, depending on the strategy chosen. As a first step towards optimal Portuguese-English SMT, in this work we further investigate several training and decoding parameter choices, and compare their outputs in a number of experiments. Whilst we do not exhaustively cover all possible alternatives, we expect to gradually overcome the complexity of the task and arrive at a fine-tuned configuration for the Portuguese/English language pair.

2. Experiments

We used Moses (Koehn et al., 2007) and 3-gram language models to develop a basic phrase-based SMT system described in Aziz et al. (2009). In order to test different combinations of training and decoding parameters, a number of experiments for Brazilian Portuguese (BP) and American English (AE) translation will be presented. All experiments made use of a Portuguese-English parallel corpus taken from the

Environment, Science, Humanities, Politics and Technology supplements of the on-line edition of the "Revista Pesquisa FAPESP"[1], a Brazilian magazine on scientific news. The training data for each experiment consisted of a set of about 17,000 sentences pairs. For testing purposes each experiment used 649 previously unseen sentence pairs. When relevant, we have also used a development set conveying another 1,989 sentences pairs.

The different parameter choices were compared by measuring BLEU (Papineni et al., 2002) and NIST (NIST, 2002) scores, two of the best-known evaluation metrics used in the MT field. Generally speaking, both BLEU and NIST scores represent the number of n-grams shared between machine and human (reference) translations. BLEU scores range from 0 to 1, whereas the maximum NIST value depends on the size of the data set. In both cases, the higher the score, the better the translation quality.

2.1. Training options

In this section we will focus on four SMT training options: the alignment heuristics, the maximum phrase-length, the use of lexical weighting and tuning. In all cases, Giza++ alignment options were set to *m1*=5, *m2*=0, *mh*=5, *m3*=3, *m4*=3, and distortion limit was left at its default value 6. For details, see Koehn et al. (2007).

Our first goal is to look into the alignment heuristics, that is, the strategy that determines how potential alignment points are connected. In our previous work we used Moses *grow-diag-final-and* (*gdfA*) heuristics, in which alignments may be established as directly to the left, right, top, or bottom (corresponding to the 'grow' step) and also diagonally (the 'grow-diag' strategy). As for the words that do not neighbour, the 'final' option adds the alignments in which at least one word is unaligned, and only alignment points between two (hence 'and') unaligned words are added (Koehn et al., 2007). Now we would like to check how the use of a less restrictive heuristics, *grow-diag-final* (*gdf*) may impact the BLEU/NIST scores. The results are shown in Table 1 below:

Table 1. Alignment heuristics on AE-BP translation.

Exp.	BLEU	NIST	Alignment	Mx_Pl	Lex_w	Tuning
1	0.3072	7.3891	gdfA	7	yes	no
2	0.3053	7.3228	gdf	7	yes	no

The *gdfA* heuristics still produced slightly superior BLEU/NIST scores, and for that reason we will keep this heuristics unchanged in our next experiment. We will now examine the effect of Moses maximum phrase-length parameter (*Mx_Pl*) on translation output. More specifically, we will vary the values of *Mx_Pl* from 7 to 3 as follows.

Table 2. Maximum phrase length in AE-BP translation.

Exp.	BLEU	NIST	Alignment	Mx_Pl	Lex_w	Tuning
1	0.3072	7.3891	gdfA	7	yes	no
3	0.3053	7.3608	gdfA	3	yes	no
4	0.3061	7.3766	gdfA	4	yes	no
5	0.3071	7.3886	gdfA	5	yes	no
6	0.3072	7.3891	gdfA	6	yes	no

From the above we observe that BLEU/NIST scores deteriorate as we attempt to use shorter maximum phrase-length, and that *Mx_Pl*=6 has exactly the same effect as

[1] http://www.revistapesquisa.fapesp.br/

Mx_Pl=7. However, in order to keep the experiment setting unchanged, we will continue to use *Mx_Pl*=7 in our next experiment, in which we examine the role of lexical weighting option (i.e., the use of phrase translation probabilities augmented with lexical translation probabilities.) To this end, Experiment 1 was re-run without lexical weighting. The results are shown in Table 3 below.

Table 3. Lexical weighting in AE-BP translation.

Exp.	BLEU	NIST	Alignment	Mx_Pl	Lex_w	Tuning
1	0.3072	7.3891	gdfA	7	yes	no
7	0.3052	7.3453	gdfA	7	no	no

The results for Experiment 7 above show that *not* using lexical weighting reduces the overall translation quality, and are consistent with the findings in Koehn et al., (2007). Thus, we will once again keep this option unchanged as in Experiment 1.

Finally, Experiment 1 was re-run using tuning applied on our development data set. This additional parameter estimation has increased our training times from round 30 minutes to 15 hours to be completed, but it did improve results as follows:

Table 4. Tuning in AE-BP translation.

Exp.	BLEU	NIST	Alignment	Mx_Pl	Lex_w	Tuning
1	0.3072	7.3891	gdfA	7	yes	no
8	0.3166	7.5349	gdfA	7	yes	yes

2.2. Decoding Options

In this section we will examine the role of distortion limits (*DL*) in the decoding process. The value of the distortion limit parameter represents the number of words that are allowed to be reordered during translation, and may range from zero (no reordering) to infinity (allowing sequences of words of arbitrary length to be reordered.) In all our previous experiments, *DL* was set to its default value 6. Now that we have found a possibly ideal combination of training parameters, we will vary the value of the *DL* parameter as an attempt to improve results even further. In doing so, we are of course aware that different *DL* values may require different training parameters for optimal results. However, given that training is a time-consuming task (particularly, when tuning is used, as in Experiment 8) we are presently unable to explore all possible interactions between these parameters. For the same reason, we will use the set of training parameters originally obtained for AE-BP translation (as seen in Experiments 1-8) also for BP-AE translation, even though the AE-BP configuration is most likely suboptimal in the opposite direction. In what follows we will test possible *DL* values ranging from 0 to 7, and the option with no distortion limit (*DL*=∞). All tests are based on the outcome of our previous Experiment 1 and 8.

Table 5. Distortion limits in AE-BP translation.

Exp 1	0	1	2	3	4	5	6	7	∞
BLEU	0.2971	0.2971	0.3059	0.3070	0.3069	0.3067	0.3072	0.3073	0.3166
NIST	7.3140	7.3140	7.3705	7.3841	7.3874	7.3848	7.3891	7.3915	7.5349

Exp 8	0	1	2	3	4	5	6	7	∞
BLEU	0.3039	0.3039	0.3187	0.3181	0.3170	0.3163	0.3166	0.3165	0.2691
NIST	7.4083	7.4083	7.5309	7.5366	7.5351	7.5310	7.5349	7.5349	7.4658

Table 6. Distortion limits in BP-AE translation.

Exp 1	0	1	2	3	4	5	6	7	∞
BLEU	0.3448	0.3448	0.3497	0.3499	0.3500	0.3499	0.3497	0.3497	0.3427
NIST	8.0107	8.0107	8.0558	8.0543	8.0532	8.0518	8.0510	8.0495	8.0304

Exp 8	0	1	2	3	4	5	6	7	∞
BLEU	0.3515	0.3515	0.3600	0.3591	0.3585	0.3582	0.3579	0.3578	0.3457
NIST	8.0992	8.0992	8.1953	8.1914	8.1831	8.1799	8.1790	8.1788	8.1246

Optimal AE-BP translation with tuning (Experiment 8) is obtained with *DL*=2 (according to BLEU) or *DL*=3 (NIST). In the opposite direction (BP-AE), the best results with tuning are obtained with *DL*=2 according to both metrics. An optimal value close to *DL*=3 is consistent with other studies in the field (e.g., Koehn et al. 2003).

3. Discussion

We have presented a series of experiments in statistical phrase-based MT applied to Portuguese/English in which a number of training and decoding parameters were tested. Given the computational costs of the training procedure, not all possible combinations could be presently examined, and a single estimate was applied to the translation tasks in both directions. As future work, we intend to investigate the interactions in Giza++ alignment models, and their relation to the present definitions for both BP-AE and AE-BP translation individually.

Acknowledgements

The authors acknowledge support by FAPESP and CNPq.

References

Aziz, Wilker Ferreira, Thiago Alexandre Salgueiro Pardo and Ivandré Paraboni (2009) "Statistical Phrase-based Machine Translation: Experiments with Brazilian Portuguese". VII Encontro Nacional de Inteligência Artificial (ENIA-2009).

Koehn, Philipp et. al. (2007) "Moses: Open Source Toolkit for Statistical Machine Translation". ACL-2007.

Koehn, Philipp; Franz Josef Och, and Daniel Marcu (2003) "Statistical phrase-based translation". HLT-NAACL-2003, pages 48-54.

NIST (2002) "Automatic Evaluation of Machine Translation Quality using n-gram Co-occurrence Statistics". *http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf*

Papineni, K.; S. Roukos; T. Ward and W. Zhu (2002) "BLEU: a Method for Automatic Evaluation of Machine Translation". ACL-2002, pages 311-318.