Statistical Analysis of Alignment Characteristics for Phrase-based Machine Translation

Patrik Lambert and Simon Petitrenaud
LIUM, University of Le Mans
Avenue Laënnec
72085 Le Mans Cedex 9, France
patrik.lambert@lium.univ-lemans.fr
simon.petit-renaud@lium.univ-lemans.fr

Yanjun Ma and Andy Way
CNGL, Dublin City University
Glasnevin
Dublin 9, Ireland
yma@computing.dcu.ie
away@computing.dcu.ie

Abstract

In most statistical machine translation (SMT) systems, bilingual segments are extracted via word alignment. However, there lacks systematic study as to what alignment characteristics can benefit MT under specific experimental settings such as the language pair or the corpus size. In this paper we produce a set of alignments by directly tuning the alignment model according to alignment F-score and BLEU score in order to investigate the alignment characteristics that are helpful in translation. We report results for a phrase-based SMT system on Chinese-to-English IWSLT data, and Spanish-to-English European Parliament data. With a statistical analysis into alignment characteristics that are correlated with BLEU score, we give alignment hints to improve BLEU score using a phrase-based SMT system and different types of corpus.

1 Introduction

Most statistical machine translation (SMT) systems (*e.g.* phrase-based, n-gram-based) build their translation models from word alignments trained in a previous stage. Many papers have shown that intrinsic alignment quality is poorly correlated with MT quality (for example (Vilar et al., 2006)). Accordingly, some research has attempted to tune the alignment directly according to specific MT evaluation metrics (Lambert et al., 2007). In this paper we instead try to discover which alignment characteristics improve or worsen translation quality by analysing the word alignment produced by the alignment model with different tuning criteria. The findings can potentially benefit our understanding of existing SMT systems as well as designing novel word alignment models.

A considerable amount of research effort has been devoted to the investigation of alignment characteristics that benefit MT. These characteristics include alignment precision and recall (Ayan and Dorr, 2006; Chen and Federico, 2006; Mariño et al., 2006; Fraser and Marcu, 2007), long-distance links (Vilar et al., 2006), unlinked words (Guzman et al., 2009; Lambert et al., 2009), etc. In most of the related papers some alignment characteristics are usually considered, and the impact on MT of alignments with different values for these characteristics is evaluated.

In this work, we start from an initial alignment and tune it directly according to an intrinsic alignment quality metric (F-score, see Section 3.4) and according to an extrinsic translation quality metric (BLEU score (Papineni et al., 2002)). In this way, we can investigate for *any* alignment characteristic how it is affected by the change of tuning criterion. If there exist alignment characteristics which are helpful in translation, they should not depend on the specific aligner used. However, they could depend on parameters such as the type of MT system, the language pair, or the corpus size or type. In this way we can study more systematically how the considered characteristics depend on these parameters. We report results for the Moses phrase-based SMT system (Koehn et al., 2007). We undertook this comparison on two different tasks: translation from Chinese to English, trained with data provided within IWSLT evaluation campaigns, and translation from Spanish to English, trained on collections of three different sizes (0.55, 2.7 and 34.6 million words) of the European Parliament proceedings. Finally, in this paper we perform a de-

tailed statistical analysis of the data, focusing on the correlations between various alignment characteristics and variables that can reflect the quality of the translation, such as BLEU score or the number of untranslated words.

The remainder of the paper is organised as follows. In Section 2, we present a list of word alignment characteristics investigated in our paper. Section 3 describes the experimental setup including the word alignment model, data and evaluation methods. In Section 4 the results are discussed and a statistical analysis into the correlation between word alignment characteristics and translation quality is conducted. Finally, we draw conclusions and point out avenues for future work.

2 Word Alignment Characteristics Investigated

To better describe word alignment characteristics, we give the following definitions.

link Association between a source word (or position) and a target word (or position). Example: 0-2.

alignment Set of links. Example: {0-0, 1-1, 2-3, 3-2, 2-4}.

cluster *Minimal* set of source and target words such that all source words are linked only to the target words in the same set, and all target words are linked only to the source words in the same set. In the former example there are 4 clusters: {0-0}, {1-1}, {2-3, 2-4} and {3-2}.

gap Embedded position between two target (source) words linked to the same source (target) word. The word at this position might be linked to other source words.

span Distance between the first and the last position of a cluster, in the source or target side

For each system we calculated the value for the following alignment and translation quantities:

Translation	
pb_notr	Number of untranslated words (words present in the training corpus but not translated)
PB	BLEU score

Alignment	
R	Recall
P	Precision
F	F-score
dist	Distortion: average difference between source and target positions of a link
crosspl	Percentage of crossing links
clen	Crossing link distortion
gaps	Number of gaps per word
span	Span per word
links	Number of links
unlnk	Number of unlinked words
Distribution of word involved in:	
unlkpc	null links (unlinked words)
1-to-1	one-to-one alignments
1-to-n	one-to-many alignments
n-to-m	many-to-many alignments
1n-to-m	any-to-many alignments (see Section 4.2.1)

3 Experimental Setup

Our aim is to obtain alignments optimised according to both an intrinsic and an extrinsic criterion. For each criterion, the optimisation consists of maximising a function of the alignment system parameters: F-score (intrinsic criterion), and BLEU score (extrinsic criterion). We use a discriminative alignment system (Moore, 2005) because of its flexibility. First we describe this aligner, and then the optimisation procedure.

3.1 Discriminative Alignment System

This alignment system (Lambert and Banchs, 2008) implements a log-linear combination of N feature functions which are calculated at the sentence pair level. The alignment is performed in two passes. First pass features include word association models based on IBM model 1 probabilities (Brown et al., 1993), an unlinked word model proportional to the IBM model 1 NULL link probability, a feature counting the number of links in the hypothesis, distortion models, etc.

In the second alignment pass, the association score model with IBM1 probabilities and the unlinked model are substituted by two improved models benefiting from the first-pass links: an association score model with relative link probabilities, and source and target fertility models giving the probability for a given word to have one, two, three or four or more links.

The best hypothesis is the one with best score for the weighted sum of feature functions. To find it, we implemented a beam-search algorithm based on dynamic programming. For a given sentence pair, the three best links for each source *and* for each target word are considered in search.

3.2 Alignment Optimisation Procedure

As already mentioned, we want to maximise a function of the alignment parameters, which for our alignment system are the weights λ_i of the feature functions. Thus, the function to be maximized is defined as $function(\lambda_1, \ldots, \lambda_N)$, where $function$ refers either to F-score or to BLEU score. The parameters of the first and second alignment passes were optimised together.

An optimisation algorithm[1] iteratively updates the alignment parameters so as to maximise the objective function. At each iteration, the corpus is aligned and either the alignment is evaluated to calculate the F-score, or an SMT system is built from the alignments and is evaluated to calculate the BLEU score.

3.3 Data

In order to track relevant alignment characteristics depending on language pair or corpus size, we conducted experiments on two distinct language pairs and different corpus sizes.

3.3.1 Spanish–English Europarl Task

The experiments were conducted using the TC-STAR OpenLab[2] Spanish–English EPPS parallel corpus, which contains proceedings of the European Parliament. We tuned our alignment system on two subsets extracted by randomly selecting 100,000 and 20,000 sentence pairs (these subsets will be referred to as 'ran100k' and 'ran20k' respectively). We built SMT systems from the optimum alignments obtained on each of these subsets. We also aligned the whole corpus (referred to as 'full') with the optimum weights obtained by tuning on the ran100k corpus, and built SMT systems from these alignments.

To calculate the F-score in alignment tuning we used freely available[3] alignment test data (Lambert et al., 2005). We divided the alignment test data into a 246-sentence development set and a 245-sentence test set. For MT evaluation, we had a development set of 735 sentences (MERT Dev, two references) for the internal SMT MERT procedure, a development set of 1008 sentences to calculate the BLEU score at each optimisation iteration (MT Dev, two references), and a test set of 1094 sentences to realise an extrinsic evaluation of the optimal alignment system (MT Test, two references).

3.3.2 Chinese–English BTEC Task

Another set of experiments was carried out using the Chinese–English data sets provided within the IWSLT 2007 evaluation campaign, extracted from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002). We also wanted to study the impact of the corpus size, but no more data were available to build a more informed SMT system; if we had taken another corpus, our BTEC-based alignment reference would have failed to evaluate the corresponding alignments. In order to simulate an easier task, we selected instead an 'easier' development set for alignment tuning by removing some sentences containing out-of-vocabulary (OOV) words. We also created an 'easier' test set with the same method.

Training data consisted of the default training set, to which we added the sets devset1, devset2. The resulting corpus contains 41.5k sentence pairs having respectively 9.4 and 8.7 words on average for English and Chinese. English and Chinese vocabulary sizes are respectively 9.8k and 11.4k.

Manual annotation of word alignment was carried out on devset3, of which 251 sentence pairs were used as the development set and 251 for testing. For MT evaluation, we used IWSLT 2006 test set (500 sentences, 6.1k words, 7 references) as the development set for the internal SMT MERT procedure. We used devset4 (489 sentences, 5.7k words, 7 references) as the development set to calculate the BLEU score at each alignment optimisation iteration. Our 'easier' development set ('DevEasy') was a subset of devset4. We tuned our alignments on both devset4 and DevEasy and compared the results. Our test set was IWSLT 2007 test set (489 sentences, 3.2k words, 6 references). and our 'easy' test set was a subset of it.

The number of OOV words in each development and test set are reported in Table 1.

[1] We used the SPSA algorithm (Spall, 1992), which is a stochastic implementation of the conjugate gradient method which requires only two evaluations of the objective function, regardless of the dimension of the optimisation problem.

[2] http://www.tcstar.org/openlab2006

[3] http://gps-tsc.upc.es/veu/LR

	MERT Dev.	MT Dev.	MT Test
full	112	41	32
rank100k	332	205	211
rank20k	787	533	530
DevEasy	163	38	22
devset4	163	118	79

Table 1: OOV words in development and test sets for the three Spanish–English tasks (full, ran100k and ran20k) and the two Chinese–English tasks (DevEasy and devset4).

3.4 Evaluation

Intrinsic (i.e. alignment) evaluation was performed with precision (P), recall (R) and F-score (F). In both tasks, the manual alignment reference contained mainly unambiguous (or Sure) links, and some possible links (respectively 33.3% and 12.9% for Spanish–English and Chinese–English references). The scores were calculated in the standard way, as shown in (1):

$$P = \frac{|\mathcal{A} \cap \mathcal{G}|}{|\mathcal{A}|}, \quad R = \frac{|\mathcal{A} \cap \mathcal{G}_\mathcal{S}|}{|\mathcal{G}_\mathcal{S}|}, \quad F = \frac{2PR}{P+R}, \tag{1}$$

where $\mathcal{A}$, $\mathcal{G}_\mathcal{S}$ and $\mathcal{G}$ are respectively the computed link set, the reference sure link set, and the total reference link set.

Extrinsic evaluation was performed with the BLEU score (Papineni et al., 2002). Translations were produced either by Moses (Koehn et al., 2007) with all default parameters, or by a baseline n-gram-based system with constrained reordered search (Crego and Mariño, 2007). In order to limit the error introduced by MERT, we ran 4 MERT instances, each with a different random seed. We then either consider the average of the 4 values, or take the 4 values into account in the statistical analysis of the results.

4 Results and Statistical Analysis

4.1 Translation Results

We produced 10 alignment sets in total obtained using different methods. This includes 3 baseline sets, corresponding to combinations of the Giza++ (Och and Ney, 2003) source–target and target–source alignments computed by Moses scripts: intersection (I), union (U) and grow-diag-final heuristic (GDF) (Koehn et al., 2003). 6 sets were produced with the optimum weights of the discriminative aligner (Section 3.2) resulting from optimisations according to F-score, to the phrase-based system BLEU score and to the n-gram-based system BLEU score (referred to as F, PB and NB, respectively). Because the optimisation algorithm can get stuck in a poor local maximum, the optimisation with each criterion was performed with three different random seeds. To have an idea of the error introduced by the optimisation process, we kept the weights of the two optimisations which reached the highest values in the development set. They are denoted with index 1 or 2 (as in F1 and F2). Finally, we also used the initial weights of the aligner to produce a set of alignments.

	full	ran100k	ran20k	DevEasy	devset4
F1	55.8	51.0	46.1	37.4	35.3
F2	56.0	51.1	46.2	37.2	35.1
NB1	55.8	51.0	46.1	38.2	34.7
NB2	55.8	50.9	45.9	37.1	35.1
PB1	56.0	51.2	46.3	37.9	35.1
PB2	56.3	51.4	46.5	38.1	35.6
I	55.6	50.7	46.0	36.1	33.8
U	56.7	51.1	46.2	35.2	33.1
GDF	56.3	51.2	46.2	35.8	34.0

Table 2: BLEU score using different alignment sets on the Spanish–English test data and Chinese–English test data

Table 2 shows the performance of the phrase-based SMT system using the 10 different alignments described above. The optimisation procedure was effective for this system. The best systems built from discriminative alignments were indeed those optimised with the phrase-based BLEU score as the objective function. When the alignment weights were tuned on the corresponding training corpus (all tasks except the 'full' corpus, for which alignments were tuned on only a 100k-sentence subset), alignments optimised according to BLEU score also yielded better phrase-based SMT systems than Giza++ combinations.

4.2 Statistical Analysis

4.2.1 Methodology

In this section, our aim is to investigate which variables are relevant for improving the results, especially in terms of BLEU score. For each task, we have a large number of variables and $n = 10$ systems using the 10 sets of alignments described in Section 4.1.

We started our analysis by a Principal Component Analysis to have a graphical overview of the relationship between the variables. Then, more precisely, we studied the correlation between the BLEU score and other variables in the different

tasks. We made correlation tests (Rodgers and Nicewander, 1988), which consist in choosing between the null hypothesis (H_0) for which there is no association between two variables X and Y, and the alternative hypothesis (H_1), for which there is an association. If we have a series of n measurements of X and Y written as $(x_i, y_i)_{i=1}^n$, then the sample correlation coefficient r_{XY} can be used to estimate the population correlation coefficient, and is defined as in (2):

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_X s_Y}, \qquad (2)$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, and s_X and s_Y are the sample standard deviations of X and Y. Let $\alpha \in]0, 1[$ be the risk of rejecting hypothesis H_0 by mistake, and $S_{1-\alpha,n}$ a threshold depending on the error risk α and the sample size n. Then if $|r_{XY}| < S_{1-\alpha,n}$ a we accept H_0, otherwise, we reject H_0. The threshold $S_{1-\alpha,n}$ for $n = 10$ systems and a risk $\alpha = 0.05$ is about 0.63.

The hypothesis testing for correlation between two random variables X and Y requires the assumption that both variables are distributed normally. We proposed to check this assumption with the goodness of fit version of the well-known Kolmogorov-Smirnov test. We made this test for each task and each variable. Since we use 10 systems, most variables pass this test. For example, in the "full" task, 3 variables out of 16 did not pass the test: span, gaps and many-to-many alignment variables. To investigate some effect of many-to-many alignments, we studied the sum of one-to-many and many-to-many alignments (called "any-to-many" alignments). This variable passes the normality test. When this assumption is checked, we can make the correlation test.

Because we have 4 BLEU score values (see Section 3.4), the value of the correlation coefficient is somewhat uncertain. In order to take this uncertainty into account, we also computed an interval of possible correlations in a Monte-Carlo way. Concretely, for each system, we select randomly an $\mathcal{N} = 10000$-sample of one of the 4 possible values of the BLEU score with a uniform distribution. Then we obtain a multivariate sample of $\mathcal{N}$ 10-sized vectors and for each variable, we can compute an $\mathcal{N}$-sample of the correlation between the BLEU score and the variable. With the empirical distribution function $\widehat{F}$ of the resulting correlation distribution, it is possible to build robust fluctuation intervals for the correlation: $[r_{\frac{\beta}{2}}, r_{1-\frac{\beta}{2}}]$ containing a proportion of $1-\beta$ of the values, with r_γ, the quantile of order γ of $\widehat{F}$.

4.2.2 Characteristics Helping to Improve BLEU Score

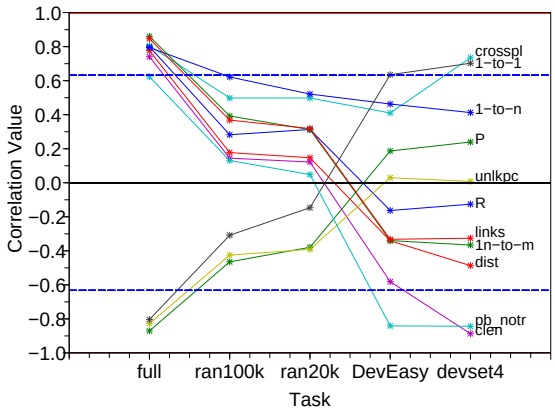

Figure 1: Correlation between the BLEU score and a number of alignment statistics, for a number of tasks: Spanish–English: full, ran100k, ran20k; Chinese–English: DevEasy and devset4. The dashed horizontal lines mark the correlation significance threshold (Section 4.2.1). The considered variables are crosspl, 1-to-1, 1-to-n, P, unlkpc, R, links, 1n-to-m, dist, pb_notr, clen (see Section 2).

Figure 1 shows the correlations between the BLEU score and most variables defined in Section 2 (the variables omitted are either redundant or do not pass the normality test). The BLEU score considered is the average of the 4 values corresponding to MERT processes with different random seeds (see Section 3.4).

Before analysing Figure 1 it is important to point out that the correlation value for most variables is significant only in the 'full' task. We nevertheless think that the trend of the correlation value versus the corpus size is interesting.

A number of variables range from negatively correlated with BLEU score to positively correlated with BLEU score depending on the task. Thus the impact on BLEU score of these variables greatly depends on the size of the corpus. Typically the correlation value is significantly positively or negatively correlated with BLEU score in the 'full' task. For the ran100k and ran20k tasks, the correlation value is decreased below the significance threshold but remains of the same sign. This means that the correlation remains positive or negative, but the confidence degree is decreased. For example if the correlation value $r_{XY} = \pm 0.4$,

we may consider that both variables are correlated with an error risk $\alpha = 0.25$. If $|r_{XY}| = 0.25$, $\alpha = 0.5$, thus there are as many chances of error as success to consider that some correlation exists. For Chinese–English tasks, this value is close to zero or of opposite sign. Two variables do not follow this trend: the percentage of words involved in one-to-many alignments, and the number of crossing links.

Disappointingly, but nonetheless interesting, no variable is significantly correlated (positively or negatively) with BLEU score for all corpora. The variable which is most consistently positively correlated with BLEU score is the percentage of words involved in one-to-many alignments, but it is above the significance threshold only for the Spanish–English full task. The number of crossing links is also always positively correlated with BLEU score, although only significantly so in the full and devset4 tasks. The variable which is always negatively correlated with BLEU score (although not always significantly) or with no correlation is the percentage of unlinked words.

The variables which are positively correlated with BLEU score in the 'full' task take higher values in dense alignments: the number of links, the ratio of words in one-to-many alignments, the alignment recall, the average link distortion, the average link crossing length or the number of untranslated words. Conversely, the variables negatively correlated with BLEU score in this task take higher values in sparse alignments: the ratio of words in one-to-one alignments, the alignment precision, the ratio of unlinked words. Thus, two clear conclusions from this correlation analysis are that with larger corpora, dense alignments are better for phrase-based SMT, while with smaller corpora, more precise, sparser alignments are required.

These findings are illustrated in Figure 2. The whole range of possible correlation values given the several BLEU score values, as explained in Section 4.2.1, is displayed on the graph. Figure 2 shows that alignment recall is rather positively correlated with BLEU score (although not necessarily significantly) with larger corpora and negatively with smaller corpora, and conversely for the alignment precision.

Figure 3 displays the correlation between the percentage of crossing links and BLEU score, and between the average distortion of crossing links and BLEU score. It seems that crossing links

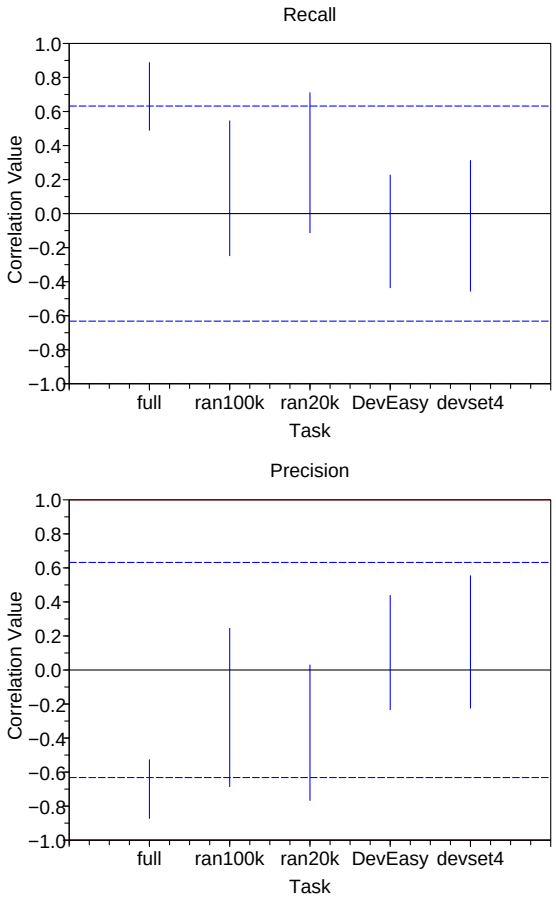

Figure 2: Correlation of BLEU score with the alignment recall (top) and with the alignment precision (bottom).

themselves are not problematic, since for all tasks the correlation interval mostly remains in the positive half of the figure. However, the smaller the corpus, the more problematic long-distance crossing links may be. Thus for small corpora, avoiding some long-distance links may improve BLEU score.

Figure 4 represents the correlation between the BLEU score and the number of untranslated words versus the task. It shows that the less information the translation model has to translate the test set, the more negative impact the number of untranslated words have on the BLEU score. For Spanish to English tasks, there seems to be no correlation or even a positive correlation for the larger corpus. For Chinese–English tasks, the correlation value ranges around the negative threshold, meaning that BLEU score may be improved by reducing the number of untranslated words. Thus it is relevant to investigate how to reduce the number of untranslated words from an alignment point of

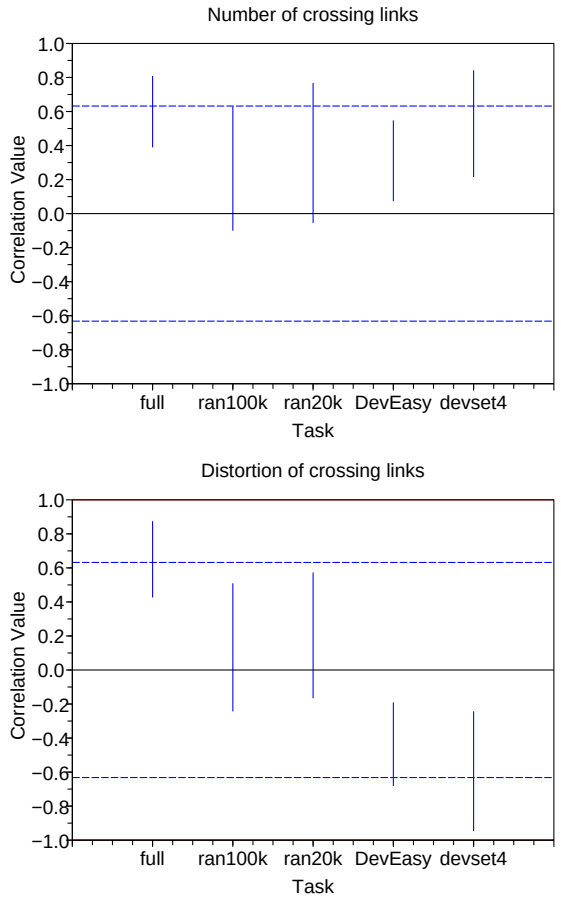

Figure 3: Correlation of BLEU score with the percentage of crossing links (top) and with the average distortion of crossing links (bottom).

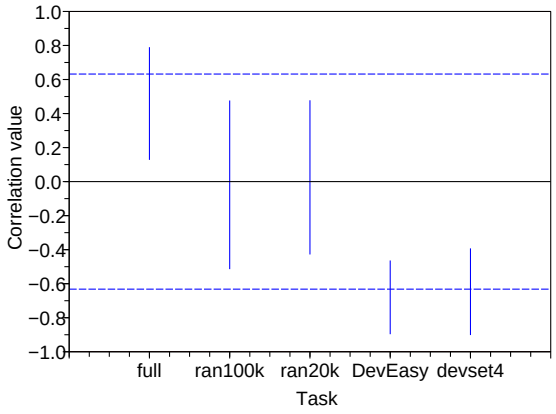

Figure 4: Correlation between the phrase-based BLEU score and the number of untranslated words.

	full	ran100k	ran20k	DevEasy	devset4
1n-to-m	0.585	0.929	0.906	0.724	0.721
links	0.541	0.932	0.897	0.711	0.704
R	0.493	0.944	0.866	0.579	0.551
dist	0.479	0.952	0.965	0.381	0.827
1-to-n	0.410	0.701	0.565	-0.036	-0.050
crosspl	0.227	0.716	0.643	-0.590	-0.663
unlnk	-0.461	-0.884	-0.812	-0.452	-0.436
P	-0.564	-0.885	-0.857	-0.582	-0.613
1-to-1	**-0.744**	**-0.957**	**-0.969**	**-0.919**	**-0.918**

Table 3: Correlation coefficient between the number of untranslated words of the phrase-based system and several alignment variables (Section 2).

view. Although we would expect that the number of untranslated words be less relevant for the DevEasy task than for the devset4 task, note that it is not the case. So we did not succeed in simulating a larger Chinese–English corpus.

Table 3 shows how the number of untranslated words is correlated with a number of alignment variables. The only variable above the significance threshold in all tasks is the number of words involved in one-to-one alignments (negatively correlated). Thus a higher percentage of one-to-one alignments helps to reduce the number of untranslated words, whatever the amount of data. This is an intuitive result, since untranslated words never constitute alone the source or target side of a bilingual phrase. This can happen if they are unlinked or if they are involved in one-to-many or many-to-many alignments. Except for the 'full' task, the percentage of words involved in one-to-many or many-to-many alignments and the number of links are significantly positively correlated with the number of untranslated words.

5 Conclusions and further work

We tracked helpful alignment characteristics for MT by tuning a discriminative alignment system according to alignment F-score and translation BLEU score (obtained with two different MT systems), and compared the resulting alignments and their impact on MT quality (evaluated with the BLEU score). We conducted this experiment on five distinct tasks, representing different corpus sizes and language pairs. We performed a statistical analysis of the data, including Principal Component Analysis, and studies of the sample correlation coefficients between a number of alignment characteristics and variables reflecting MT quality such as the number of untranslated words or the BLEU score.

We found that for small tasks like the Chinese–English IWSLT tasks, limiting the number of untranslated words may improve BLEU score. The number of untranslated words can be reduced via a

higher percentage of one-to-one alignments, whatever the amount of data. We found that for most tasks no variable is highly correlated with BLEU score, although for the largest task correlation coefficients are higher. We were nevertheless able to draw general conclusions from the correlation analysis. With larger corpora, dense alignments are required while with smaller corpora, more precise, sparser alignments are better for phrase-based SMT. Crossing links themselves do not seem to be problematic, but avoiding some long-distance crossing links may improve BLEU score when using small corpora. Our main conclusion is that the alignment characteristics which help in translation greatly depend on the corpus size.

Acknowledgements This work has been partially funded by the European Union under the EuroMatrix Plus project (http://www.euromatrixplus.net, IST-2007.2.2-FP7-231720) and by Science Foundation Ireland (www.sfi.ie, Grant 07/CE/I1142).

References

Ayan, N. F. and B. J. Dorr. 2006. Going beyond AER: An extensive analysis of word alignments and their impact on MT. In *Proc. of Meeting of the Assoc. for Computational Linguistics*, pages 9–16, Sydney, Australia.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chen, B. and M. Federico. 2006. Improving phrase-based statistical translation through combination of word alignment. In *Proc. of FinTAL - Int. Conf. on Natural Language Processing*, Turku, Finland.

Crego, J. M. and J. B. Mariño. 2007. Improving SMT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Fraser, A. and D. Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Guzman, F., Q. Gao, and S. Vogel. 2009. Reassessment of the role of phrase extraction in pbsmt. In *Proc. of Machine Translation Summit XII*, pages 49–56, Ottawa, Canada.

Koehn, P., F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of Meeting of the Assoc. for Computational Linguistics*, pages 48–54, Edmonton, Canada.

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

Lambert, Patrik and Rafael E. Banchs. 2008. Word association models and search strategies for discriminative word alignment. In *Proc. of the Conference of the European Association for Machine Translation*, pages 97–103, Hamburg, Germany.

Lambert, P., A. de Gispert, R. E. Banchs, and J. B. Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285.

Lambert, P., R. E. Banchs, and J. M. Crego. 2007. Discriminative alignment training without annotated data for machine translation. In *Proc. of the Human Language Technology Conference of the NAACL (Short Papers)*, pages 85–88, Rochester, NY, USA.

Lambert, P., Y. Ma, S. Ozdowska, and A. Way. 2009. Tracking relevant alignment characteristics for machine translation. In *Proc. of Machine Translation Summit XII*, pages 268–275, Ottawa, Canada.

Mariño, J. B., R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A.R. Fonollosa, and M. R. Costa-jussà. 2006. N-gram Based machine translation. *Computational Linguistics*, 32(4):527–549.

Moore, R. C. 2005. A discriminative framework for bilingual word alignment. In *Proc. of Conf. on Empirical Methods in Natural Language Processing*, pages 81–88, Vancouver, Canada.

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

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of Meeting of the Assoc. for Computational Linguistics*, pages 311–318, Philadelphia.

Rodgers, J. L. and W. A. Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, February.

Spall, J. C. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Control*, 37:332–341.

Takezawa, T., E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of Third Int. Conf. on Language Resources and Evaluation 2002*, pages 147–152, Las Palmas, Spain.

Vilar, D., M. Popovic, and H. Ney. 2006. AER: Do we need to "improve" our alignments? In *Proc. of the Int. Workshop on Spoken Language Translation*, pages 205–212, Kyoto, Japan.