Tracking Relevant Alignment Characteristics for Machine Translation

Patrik Lambert, Yanjun Ma, Sylwia Ozdowska and Andy Way
School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland
{plambert,yma,sozdowska,away}@computing.dcu.ie

Abstract

In most statistical machine translation (SMT) systems, bilingual segments are extracted via word alignment. In this paper we compare alignments tuned directly according to alignment F-score and BLEU score in order to investigate the alignment characteristics that are helpful in translation. We report results for two different SMT systems (a phrase-based and an n-gram-based system) on Chinese to English IWSLT data, and Spanish to English European Parliament data. We give alignment hints to improve BLEU score, depending on the SMT system used and the type of corpus.

1 Introduction

Most statistical machine translation (SMT) systems (*e.g.* phrase-based, n-gram-based) extract their translation models from word alignment trained in a previous stage. Many papers have shown that alignment quality is poorly correlated with MT quality (for example Vilar *et al.* (2006)). Then, we can tune the alignment directly according to MT metrics (Lambert et al., 2007). In this paper we rather try to find out which alignment characteristics help or worsen translation.

In the related papers (see next section) some alignment characteristics are usually considered, and the impact on MT of alignments with different values for these characteristics is evaluated. The contributions of this paper are twofold. Firstly, the problem is considered from the inverse point of view: we start from an initial alignment and tune it directly according to a translation quality metric (BLEU score (Papineni et al., 2002)) and according to an alignment quality metric (F-score, see Section 4.3). 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 aligner used. However, they could depend on the MT system, the language pair, or the corpus size or type. The second contribution of this paper is to study more systematically how the considered characteristics depend on these parameters. We report results for two different SMT systems: a phrase-based system (Moses (Koehn et al., 2007)) and an n-gram-based system (Crego and Mariño, 2007). We performed this comparison on two different tasks: translation from Chinese to English, trained with IWSLT data (BTEC corpus, a small corpus in the travelling domain), and translation from Spanish to English, trained on a 2.7 million word corpus of the European Parliament proceedings.

First we discuss related work. In Section 3, we describe the alignment optimisation procedures according to F-score and BLEU, and give more details on the alignment system used. Then in Section 4, we provide a summary of the experiments performed on each task, together with a description of the data used. In Section 5, the results are discussed. Finally, some conclusions are provided together with avenues for further research.

2 Related Work

In this section, we review some alignment characteristics that have been observed to have some impact in phrase extraction and MT output, for the SMT approaches we consider in this paper. We will thus consider these characteristics (and more) to investigate what kind of alignment helps depending on the system and type or amount of training data.

In several papers the impact of higher precision or higher recall alignments has been studied. Ayan

and Dorr (2006) and Chen and Federico (2006) observed that higher precision alignments favoured a phrase-based SMT system. In the former case, it was observed with an English Chinese training corpus of 1.1 million running words (for the English side), and with an English Arabic corpus of 3.3 million words of News and treebank data. In the latter case, the BTEC corpus was used (Chinese, Japanese, Arabic and Italian to English, with 180k English words).

Fraser and Marcu (2007) compared the performance of translation systems trained on several alignments of varying quality. Their results on large corpora do not confirm the hypothesis that higher precision alignments help phrase-based SMT systems more than higher recall alignments. For example, among their 3 systems trained on a 67 million word French English corpus, the highest precision alignment has the best BLEU score when alignment quality is low, and the highest recall alignment is the best when alignment quality is high. Their results suggest that when there is not enough data to produce good quality alignments, increasing the alignment precision improves phrase-based SMT systems. However, with larger corpora, higher recall alignments seem to be better. Our experiments give some more insight on this point.

Mariño *et al.* (2006) observed that a higher recall alignment improved the performance of an n-gram-based translation model on the Europarl corpus (35 million running words).

Another important issue in the extraction of bilingual segments is the presence of long-distance links. Vilar *et al.* (2006) improved the translation quality of a German English phrase-based SMT system by deleting links between the English verb and the German particle part of the verb, which is situated far from the main part of the verb and produces a long-distance link.

This issue is particularly relevant for the n-gram-based system, where a unique segmentation of the sentence pair is performed. It is nevertheless partially addressed by the reordering strategy, which is based on a monotonisation of the alignment prior to the extraction of bilingual phrases. In this monotonisation process, the source words are reordered. Thus long-distance links are only problematic when a source word is linked to two or more non-adjacent target words. In this case, the possible bilingual units involving the words embedded between these target word positions cannot be extracted.

Long-distance links can also restrict the number of bilingual phrases which are extracted in Moses. In this case, the bilingual phrases involved with embedded words are still extracted. However, those bilingual phrases involving the long-distance one-to-many link itself may be large and thus not easy to reuse. The same problem may happen with long crossing links.

Thus we expect that alignments optimised according to BLEU will have shorter links, or shorter crossing links, or fewer embedded words than manual alignments.

3 Alignment Optimisation Procedure

Our aim was to obtain alignments optimised according to both an intrinsic and an extrinsic criterion. To achieve this, we used a discriminative alignment system (Moore, 2005) because of its flexibility. For both criteria, the optimisation consisted of maximising a function of the alignment system parameters: F-score (intrinsic criterion) and BLEU score (extrinsic criterion). First we describe the alignment system used, then the optimisation procedure.

3.1 Discriminative Alignment System

This aligner implements a log-linear combination of feature functions calculated at the sentence pair level. In a first pass, the training corpus was aligned selecting for each sentence pair $(\mathbf{s}, \mathbf{t})$ the alignment hypothesis $\hat{\mathbf{a}}$ which maximises a combination of various models, as expressed in (1):

$$\hat{\mathbf{a}}^{(1)} = \arg\max_{\mathbf{a}} \lambda_{a1}^{(1)} h_{a1} + \lambda_{a2}^{(1)} h_{a2} + \lambda_{lb}^{(1)} h_{lb} + \lambda_{um}^{(1)} h_{um} + \lambda_{cn}^{(1)} h_{cn} + \lambda_{cl}^{(1)} h_{cl} + \lambda_{hp}^{(1)} h_{hp} \quad (1)$$

where h stands for the feature functions $h(\mathbf{s}, \mathbf{t})$ used, and the λs are their corresponding weights. h_{a1} and h_{a2} are word association models based on source-target and target-source IBM model 1 probabilities (Brown et al., 1993). h_{lb} is proportional to the number of links in $\mathbf{a}$. h_{um} is an unlinked word model proportional to the IBM model 1 NULL link probability. h_{cn} and h_{cl} are distortion models, counting respectively the number and amplitude (the difference between target word positions) of crossing links. Finally, h_{hp} is a "hole penalty" model,

proportional to the number of embedded positions between two target words linked to the same source words (or vice-versa).

We performed a second alignment pass in which the association score model with IBM1 probabilities and the unlinked model were substituted by two improved models benefiting from the first-pass links: an *A*ssociation score model h_{ar} with *R*elative link probabilities (Melamed, 2000), and source and target fertility models (h_{fs} and h_{ft}) giving the probability for a given word to have one, two, three or four or more links. Second pass models are listed in (2).

$$\hat{\mathbf{a}}^{(2)} = \arg\max_{\mathbf{a}} \lambda_{ar}^{(2)} h_{ar} + \lambda_{lb}^{(2)} h_{lb} + \lambda_{fs}^{(2)} h_{fs} + \lambda_{ft}^{(2)} h_{ft} + \lambda_{cn}^{(2)} h_{cn} + \lambda_{cl}^{(2)} h_{cl} + \lambda_{hp}^{(2)} h_{hp} \quad (2)$$

To find the best hypothesis, we implemented a beam-search algorithm based on dynamic programming. In a given sentence pair, the best 3 links for each source *and* for each target word are considered in search.

The parameters of the first and second alignment passes were optimised together, to give the following objective function:[1]

$$FUNCTION(\lambda_{a2}^{(1)}, \lambda_{lb}^{(1)}, \lambda_{um}^{(1)} h, \lambda_{cn}^{(1)}, \lambda_{cl}^{(1)}, \lambda_{hp}^{(1)}, \lambda_{lb}^{(2)}, \lambda_{fs}^{(2)}, \lambda_{ft}^{(2)}, \lambda_{cn}^{(2)}, \lambda_{cl}^{(2)}, \lambda_{hp}^{(2)}),$$

where $FUNCTION$ refers either to F or to $BLEU$. With this many parameters, an optimisation algorithm was necessary (see Section 3.3).

3.2 Optimisation Set-up

As mentioned above, the following objective functions were maximised:

$$F(\{aligner\ parameters\}) \quad (3)$$

$$BLEU(\{aligner\ parameters\}) \quad (4)$$

In the case of Function (3), the whole training corpus was aligned for the first pass (Equation 1). For the second pass, only manually aligned development data were aligned to calculate the F-score (see Section 4). This constitutes the first iteration of the optimisation algorithm, realised with initial parameters. Then, alignment system parameters were simply adjusted by the optimisation algorithm so as to maximise the F-score.

In the case of Function (4), the training corpus was aligned with initial parameters and these alignments were used to build either an n-gram-based or a phrase-based SMT system. The model weights were tuned via MERT (Och, 2003), with the Moses MERT utility (which was adapted to the n-gram-based system). Then a translation of a development corpus was obtained and evaluated using BLEU (see Section 4). Thus, at each iteration, the considered parallel corpus was aligned (with the two successive passes), an SMT system was built from the resulting alignments (including bilingual phrase extraction, model(s) estimation and MERT) and the development set was translated to obtain the BLEU score. At the end of this process we obtained the alignment parameters which maximise the BLEU score. Note that we used two developments sets: one for the alignment weight optimisation, one for MERT.

3.3 Optimisation Algorithm

The optimisation procedure was performed using the SPSA algorithm (Spall, 1992). SPSA 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. The SPSA procedure is in the general recursive stochastic approximation form:

$$\hat{\lambda}_{k+1} = \hat{\lambda}_k - \mathbf{a}_k \hat{\mathbf{g}}_k(\hat{\lambda}_k) \quad (5)$$

where $\hat{\mathbf{g}}_k(\hat{\lambda}_k)$ is the estimate of the gradient $\mathbf{g}(\lambda) \equiv \partial E / \partial \lambda$ at the iterate $\hat{\lambda}_k$ based on the previous evaluations of the objective function. a_k denotes a positive number that usually decreases as k increases.

We performed about 80 evaluations of the objective function. Note that in general, SPSA converges to a local maximum.

4 Experiments

4.1 Chinese English BTEC Task

The experiments were carried out using the Chinese–English datasets provided within the IWSLT 2007 evaluation campaign, extracted from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002). Training data

[1] The weights in each pass can be normalised such that one weight is set to 1. This is why $\lambda_{a1}^{(1)}$ and $\lambda_{ar}^{(2)}$ were not free parameters.

consisted of the default training set, to which we added the sets devset1, devset2 and devset3. 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 development set for the internal SMT MERT procedure, and devset4 (489 sentences, 5.7k words, 7 references) as development set to calculate the BLEU score at each optimisation iteration. Our test set was IWSLT 2007 test set (489 sentences, 3.2k words, 6 references).

4.2 Spanish English Europarl Task

Another set of experiments was conducted using a part of the TC-STAR OpenLab[2] Spanish English EPPS parallel corpus, which contains proceedings of the European Parliament. We randomly selected 100k sentence pairs, having respectively 27.2 and 28.4 words on average for English and Spanish. English and Spanish vocabulary sizes are respectively 38k and 55k words.

To calculate F-score we used freely available[3] alignment test data (Lambert et al., 2005).We randomly 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 for the internal SMT MERT procedure, a development set of 1008 sentences to calculate the BLEU score at each optimisation iteration, and a test set of 1094 sentences to realise an extrinsic evaluation of the optimal alignment system. All three sets had two references.

4.3 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 following way:

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

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 computed 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).

5 Results

We present intrinsic and extrinsic evaluation results as well as some statistics for 9 alignment sets. 3 sets are baseline sets, and correspond 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). The other sets were aligned with the optimum weights of the discriminative aligner (Section 3.1) resulting from optimisations according to the 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).

5.1 Intrinsic and Extrinsic Evaluations

Table 1 shows the evaluation results for the different computed alignments on the Chinese English and Spanish English tasks.

First, note that the optimisation procedure was effective since for each score the best systems built from discriminative alignments were those optimised with this score as objective. Note also that although the discriminative aligner could not achieve better alignments in terms of F-score than Giza++,

[2] http://www.tc-star.org/openlab2006

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

	R	P	F	BLEU NB	BLEU PB
Chinese English					
F1	79.1	87.0	82.9	30.5	35.3
F2	77.2	90.9	83.5	33.0	35.1
NB1	79.3	85.5	82.3	33.2	34.3
NB2	79.0	86.2	82.5	34.4	34.6
PB1	79.0	85.4	82.1	31.5	35.5
PB2	78.5	87.2	82.7	32.8	35.9
I	63.4	97.5	76.8	28.9	34.0
U	87.5	78.2	82.6	29.1	33.2
GDF	86.9	79.9	83.2	29.9	33.5
Spanish English					
F1	69.9	93.9	80.1	50.6	50.9
F2	69.0	94.7	79.8	50.5	51.0
NB1	68.8	91.3	78.5	50.6	50.9
NB2	68.8	93.1	79.1	50.6	51.1
PB1	70.1	90.0	78.9	50.3	51.2
PB2	70.7	89.6	79.0	50.0	51.6
I	68.8	97.1	80.5	50.2	50.7
U	81.4	78.6	80.0	50.6	51.1
GDF	80.3	80.9	80.6	50.8	51.0

Table 1: Recall (R), Precision (P), F-score (F), n-gram-based (NB) and phrase-based (PB) system BLEU scores for the different alignment sets on the Chinese English and Spanish English test data.

it was able to produce alignments resulting in better MT systems than with Giza++ combinations, except for the Spanish English n-gram system.

On Chinese English data, the impact on recall or precision of optimising alignment according to the phrase-based system BLEU score is not clear. For the n-gram based system, the effect is a decrease of alignment precision. For the Spanish English task, recall is slightly better for systems PB1 and PB2, and precision is lower for both NB and PB systems. Except for the Chinese English phrase-based system, the main effect of tuning alignment according to BLEU score seems therefore to be a decrease in precision. This suggest that in those cases, alignment precision is less relevant when the end-product is MT than when it is word alignment itself.

The MT evaluation reveals that the phrase-based system is fairly robust across alignment variations on this type of corpora. The variation in BLEU is 8% relative on IWSLT data, and only 1.8% relative on the Europarl data. The n-gram-based system is more sensitive to word alignment differences on IWSLT data, but not on the Europarl data (17% and 1.6% relative variation respectively).

Finally, we observed that the discriminative aligner has some difficulty to produce high recall alignments.

In the next sections we analyse the impact of the word alignment differences in the phrase table of the phrase-based and n-gram-based systems.

5.2 Moses Phrase Table Analysis

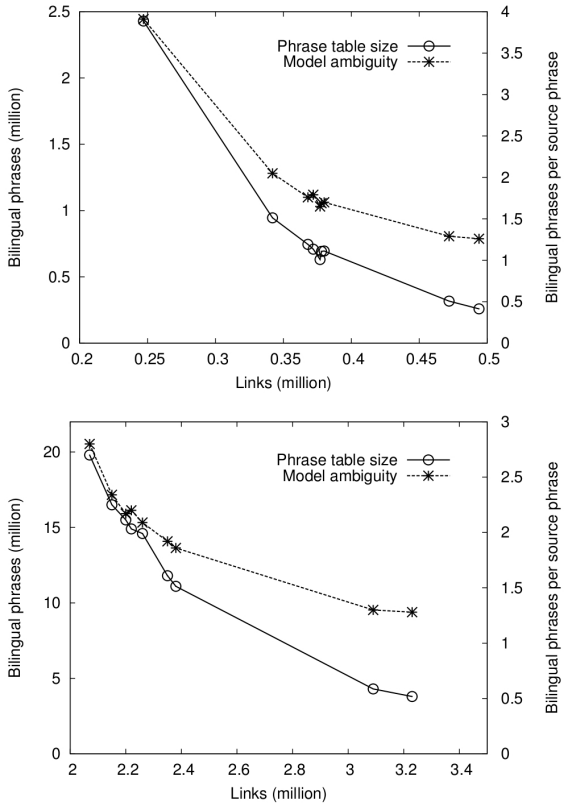

Figure 1: Number of bilingual phrases (left y axis) and number of bilingual phrases per source phrase ("model ambiguity", right y axis) versus the number of links, for Chinese English (above) and Spanish English (below).

Figure 1 shows a clear relation, on both tasks, between the phrase table size, the ambiguity of the phrase table (number of bilingual phrases per source phrase) and the number of links. The more links, the less bilingual phrases, and the less ambiguous is the model. Thus higher precision alignments will increase the coverage of the phrase-based system and

will be most helpful when little training data is available. Higher recall alignments will produce less ambiguous and thus more accurate models. Therefore, this type of alignment will be useful when enough data is available so that the coverage is not the main issue. This is in agreement with the hypothesis proposed in Section 2.

This hypothesis is further confirmed by the results depicted in Figure 2, in which the BLEU score is plotted versus the number of untranslated words (words of the translation output which were not seen in the target training corpus but were seen in the source training corpus). For IWSLT data, the less untranslated words, the higher the BLEU score. Therefore, with this small corpus the coverage is the main way of improving translation quality, and increasing alignment precision will often result in higher BLEU scores. The story is actually not so simple since there are other parameters as precision involved. Giza++ intersection is for example the alignment with highest precision, but yielded more untranslated words (and a lower BLEU score) than the discriminative alignments.

For the Spanish English task, there is no clear relation between BLEU score and the number of untranslated words. This suggests that with this amount of data the coverage is not the main issue any more. In this case one alignment characteristic which helps increasing the accuracy of the SMT model is the recall. The highest recall alignment for each aligner (U and PB2) yielded indeed the best BLEU scores (although not directly depending on the recall value). In Section 5.4 we investigate some other alignment characteristics which may be useful on this type of task.

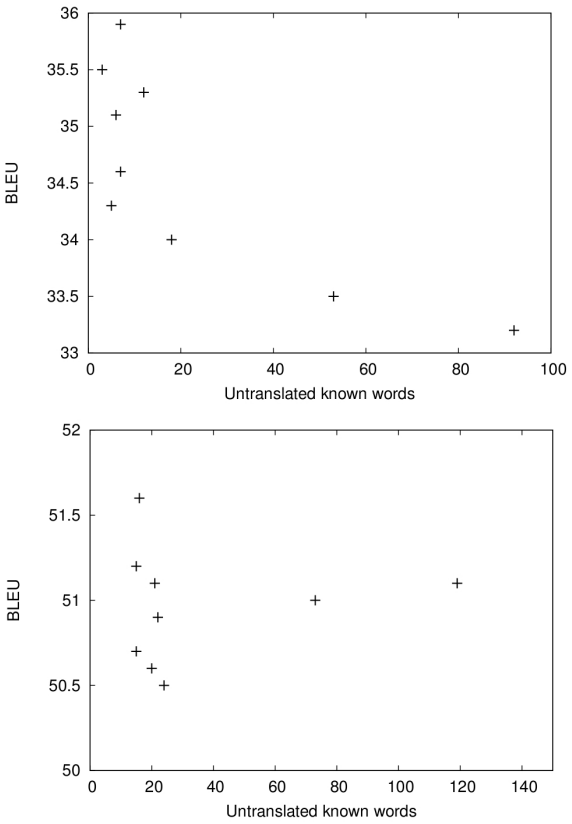

Figure 2: BLEU score versus the number of untranslated words for the Chinese English task (above) and the Spanish English task (below).

5.3 N-gram-based Phrase Table Analysis

Although we cannot display all curves because of space limitation, the BLEU score versus number of untranslated words relation for the n-gram-based system is similar to the one of the phrase-based system. On the small Chinese English corpus, the less untranslated words, the higher BLEU score. On the larger Spanish English corpus, there is no apparent relation between the two quantities. Thus the coverage is still the main problem on IWSLT data for the n-gram-based approach, whereas model accuracy is probably more important when more data is available. A difference with respect to the phrase-based approach is how more coverage is achieved. In the phrase-based approach, the bilingual phrase vocabulary increases as the number of links decreases (Figure 1). With the n-gram-based system, the relation is reversed, as depicted in Figure 3 for the Chinese English corpus (we observe a similar behaviour on Spanish English data). Except for Giza++ Union (rightmost point, at the bottom corner), the more links, the larger the bilingual units vocabulary. One possible explanation to this is that bilingual phrases with no target word (target nulled phrases) are allowed in the n-gram-based model. When the number of links increases, target nulled phrases are replaced by target phrases with words and the bilingual phrase vocabulary is enriched. As in the phrase-based approach, the model ambiguity is also reduced as the number of links increases. Thus we do not have in this case a trade-off between coverage and accuracy. As a result, higher recall alignments may

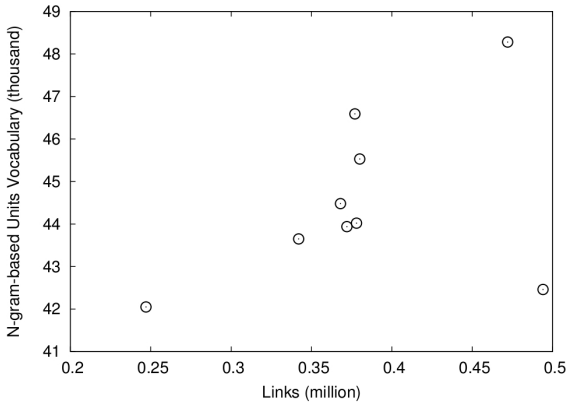

Figure 3: Number of bilingual phrases of the n-gram-based system versus the number of links for the Chinese English task.

be better than higher precision alignments, even for small corpora. This is the case for Giza++ union and GDF on the Chinese English task, which yielded a higher BLEU score than the intersection. However F1 has a higher recall than F2, but a much lower BLEU score.

5.4 Link Length and Distortion Statistics

				Embedded	
	len.	cross.	dist.	src	trg
test	4.13	3.18	2.98	2.87	2.51
F1	3.75	2.54	1.89	0.08	0.08
F2	3.74	2.35	1.87	0.06	0.20
NB1	3.75	2.40	1.70	0.07	0.21
NB2	3.69	2.52	1.42	0.05	0.11
PB1	3.86	3.10	2.22	0.07	0.29
PB2	3.83	3.41	2.14	0.05	0.98
I	3.91	2.50	3.02	0	0
U	4.96	5.37	6.50	27.29	20.94
GDF	4.78	5.15	6.02	18.18	9.73

Table 2: Average link length (len.), number of crossing links per sentence (cross.), average crossing length *i.e.* distortion (dist.), and number of source and target embedded words per sentence in each alignment set (Spanish English task).

Table 2 shows the average link length, the number of crossing links, the distortion and the number of source and target embedded words (see Section 2) per sentence, on the Spanish English task. The length of a link is defined as the absolute value of the link target and source word position difference: $length = \|position_{source} - position_{target}\|$. To calculate the number of crossing links, we sort the links according to the source and target positions as first and second keys. When the target position of a link is less than that of the anterior link according to that order, and the source position is different, we count a crossing link. The crossing (*i.e.* distortion) amplitude is the difference between both links target positions.

Links of the PB and NB alignments are shorter than in the test data. This might be due to the difficulties caused by long-distance links to the considered MT systems. However they are not shorter than links in F alignments. The alignment system might indeed discard long-distance links to gain precision.

The number of crossing links is at least as high in PB1 and PB2 systems than in the test data. However, the length of crossing links is clearly lower. This suggests that the phrase-based system BLEU score can be improved by avoiding too many long-distance crossing links. The average distortion is actually even lower for NB1 and NB2 alignments. Therefore, removing well selected long-distance crossing links may be another alignment clue to improve SMT systems.

The number of embedded words is much lower in the discriminative alignments, probably due to the difficulty of our aligner to produce high recall alignments. The n-gram-based system was expected to avoid target embedded words. This is observed if we compare NB and PB systems, but we don't understand why there are less source embedded words than target embedded words in the discriminative alignments.

6 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, and compared the resulting alignments and their impact on MT quality (evaluated with the BLEU score). We conducted this experiment for two SMT systems and on two distinct tasks, representing different corpus sizes and language pairs.

Our conclusion is that the alignment characteristics which help in translation greatly depend on the

MT system and on the corpus size.

Some related work and our results suggest that with small corpora, the coverage is the main issue governing translation quality. In the phrase-based system, coverage may be increased by increasing the alignment precision. Thus higher precision alignments may yield better SMT systems. In the n-gram-based system, higher recall alignments may still be better than higher precision ones, even on small corpora. Experiments on a fraction only of our Spanish English data would be an interesting future experiment to confirm these statements.

When more data are available, higher recall alignments allow to build less ambiguous and more accurate models. Scaling these experiments to larger corpora, as suggested above, would also be interesting to confirm this. Our results on the Spanish English data set also suggest that the phrase-based system may be improved if the alignment is more monotonic, that is if links are shorter and there is less distortion. This might also depend on the language pair, so we could repeat the experiment on a Chinese English corpus of similar size as our Spanish English corpus.

Acknowledgements

This research is supported by the Science Foundation Ireland (www.sfi.ie) (Grants 05/IN/1732 and 07/CE/I1142) and was carried out using the Irish Center for High-End Computing resources.

References

N. F. Ayan 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.

P. F. Brown, 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.

B. Chen 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.

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

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

P. Koehn, 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.

P. Koehn, 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.

P. Lambert, 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.

P. Lambert, 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.

J. B. Mariño, 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.

I. D. Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

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

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

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of Meeting of the Assoc. for Computational Linguistics*, pages 160–167.

K. Papineni, 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.

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

T. Takezawa, 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, Canary Islands, Spain.

D. Vilar, 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.