Conditional Significance Pruning: Discarding More of Huge Phrase Tables

J Howard Johnson
Interactive Information Group
National Research Council Canada
Ottawa, Ontario, Canada
Howard.Johnson@nrc-cnrc.gc.ca

Abstract

The technique of pruning phrase tables that are used for statistical machine translation (SMT) can achieve substantial reductions in bulk and improve translation quality, especially for very large corpora such at the Giga-FrEn. This can be further improved by conditioning each significance test on other phrase pair co-occurrence counts resulting in an additional reduction in size and increase in BLEU score. A series of experiments using Moses and the WMT11 corpora for French to English have been performed to quantify the improvement. By adhering strictly to the recommendations for the WMT11 baseline system, a strong reproducible research baseline was employed.

1 Introduction

Phrase-based statistical machine translation (PBSMT) (Koehn et al., 2003) requires a table of phrase pairs that indicate possible translations of sequences of words in a source language to equivalent sequences in the target language. These tables are produced from a large corpus of aligned sentences using techniques that have become well established, although improvements continue to be made.

The preferred techniques use one of the IBM models (Brown et al., 1993) that learn word alignments from a parallel corpus and from them induce phrase alignments and hence phrase pairs. Although quite successful at producing good candidate phrase pairs, erroneous phrase pairs also occur. Translation quality can be improved if these poorer phrase pairs are appropriately down-weighted (Foster et al., 2006).

However, the large bulk of typical phrase tables can cause SMT systems to require more memory and time to achieve their goal, inhibiting the use of SMT on devices with smaller resources. It also inhibits the use of phrase tables for non-PBSMT applications where quality translation equivalents are needed and techniques like smoothing are not available.

It is a matter of principle that larger and more elaborate models are always preferred over simpler ones since larger models contain the smaller ones as special cases. However, this truism is often false in practise because compensations have to be made to yield results in a timely manner. Thus as corpora have become larger, there has been a growing interest in pruning phrase tables.

In this paper, we will confirm that significance pruning (Johnson et al., 2007) is effective on huge phrase tables (up to a billion phrase pairs), and can be further improved by a relatively simple conditioning technique that is relatively cheap in resources to implement. It also confirms that for smaller phrase tables, there is room for improvement in this technique as mentioned by several other authors. The contributions of this paper are:

(1) Establish that significance pruning still provides the benefit of improved translation quality (in terms of BLEU) for large corpora; this is not the case on medium or smaller corpora where a loss is sustained. Although this is replication of known work, it needs to be revisited as phrase-table production methods evolve. Since this is a claim about exist-

ing practise, this is done on the strong community baseline of Moses and WMT11 corpora and benchmarks.

(2) A variant of significance pruning, here called *conditional significance pruning*, considering some of the inter-phrase-pair relationships is introduced. Many poor phrase pairs that are components of non-compositional phrase equivalences but are otherwise uncommon are removed. This results in a further improvement in translation quality for given phrase table size. It also removes many problematic but otherwise highly significant phrase pairs.

(3) Conditional significance pruning is evaluated in two different scenarios, one of which approximates the corpus as an extension of the phrase table. This doesn't harm the translation quality in spite of forcing the deletion of all phrase pairs that have a co-occurrence count of 1. There is strong evidence that this may be a good idea for large corpora although it is known to cause problems for smaller ones.

(4) Pruning is normally recommended as an efficiency strategy for small or low-resourced environments. This work shows that, with current phrase-table-generation methods 80% of the huge tables can be removed with an improvement to BLEU.

There is a growing literature on pruning phrase tables. Many of these focus on avoiding too much loss of translation quality. This work shows, that, at least for huge phrase tables and standard techniques, pruning is accompanied by a gain in BLEU. It is understood that, any BLEU-gain from pruning could be converted into a smoothing improvement that is better overall.

Zens et all (2012) present a technique for pruning phrase tables that directly removes the redundancy after smoothing, de-coupling the impact of harmful phrase pairs from the reduction in model size. With state-of-the-art smoothing techinques, this will be the baseline for future work.

Section 2 provides a brief overview of related work, with a discussion of other techniques for improving phrase pair quality, or applications of high quality phrase pairs other than PBSMT. Section 3 describes the technique with an example of a problem phrase pair that can be analyzed more accurately through conditioning and how its significance level can be adjusted. The remainder of the paper describes a series of experiments that use a standard baseline (WMT11 Moses baseline) on publicly available corpora (WMT11 French English) to demonstrate the effectiveness of these techniques. Section 4 describes the experimental setup, Section 5 shows the results and the paper ends with conclusions in Section 6.

2 Related Work

This work continues and builds on the work of Johnson et al (2007), Moore (2004) and Yang and Zheng (2009). Tomeh and others discuss the idea of a more complete significance test for phrase tables (Tomeh et al., 2009). While very interesting, it is not easy to implement. Other methods are discussed that use other sources of information (such as syntactic or triangulation with a third language) to help decide with phrase pairs should be removed (Huang and Xiang, 2010), (Duan et al., 2011), (Sanchis-Trilles et al., 2011), (Yu Chen et al., 2009).

3 The method of significance pruning

Following Johnson et al, (2007) the method of significance pruning retains phrase pairs that occur often in the underlying corpus. In particular, the number of times (co-occurrence count) that the source phrase occurs in source sentences and the target phrase occurs in the matched target sentence should be high enough to distinguish itself from the background of noise that occurs as a result of many types of errors that collectively can be modelled as independence. Although insufficient by itself, Johnson et al show that combined with standard phrase table generation techniques, it worked quite well. It was shown that up to 90% of the phrase table could be discarded in this way.

For this purpose, it is convenient to construct a two by two contingency table that tabulates the sentence pairs where the two types of matches occur and do not occur. Then a test of significance can be done to assess whether the degree of co-occurrence could plausibly have occurred by chance as a result of errors.

3.1 Significance testing using two by two contingency tables

Each phrase pair can be thought of as an n, m-gram $(\tilde{s}, \tilde{t})$ where $\tilde{s}$ is an n-gram from the source side of the corpus and $\tilde{t}$ is an m-gram from the target side of the corpus.

We then define: $C(\tilde{s},\tilde{t})$ as the number of parallel sentences that contain one or more occurrences of $\tilde{s}$ on the source side and $\tilde{t}$ on the target side; $C(\tilde{s})$ the number of parallel sentences that contain one or more occurrences of $\tilde{s}$ on the source side; and $C(\tilde{t})$ the number of parallel sentences that contain one or more occurrences of $\tilde{t}$ on the target side. Together with N, the number of parallel sentences, we have enough information to draw up a two by two contingency table (Table 1) representing the relationship between $\tilde{s}$ and $\tilde{t}$. The x's are defined so that the table adds up: $C(\tilde{s},\tilde{t}) + x_1 = C(\tilde{s})$, $C(\tilde{s},\tilde{t}) + x_2 = C(\tilde{t})$, $C(\tilde{s}) + x_4 = N$, $C(\tilde{t}) + x_5 = N$, and $x_2 + x_3 = x_4$.

	$\tilde{t}$	$\tilde{t}'$	
$\tilde{s}$	$C(\tilde{s},\tilde{t})$	x_1	$C(\tilde{s})$
$\tilde{s}'$	x_2	x_3	x_4
	$C(\tilde{t})$	x_5	N

Table 1: Contingency table $CT(\tilde{s},\tilde{t})$ for $\tilde{s}$ and $\tilde{t}$. (The row labeled $\tilde{s}$ shows how the $C(\tilde{s})$ sentences with the source side containing $\tilde{s}$ split into those containing $\tilde{t}$ ($C(\tilde{s},\tilde{t})$) versus those that do not. Similarly, for the column labeled $\tilde{t}$. $\tilde{t}'$ means "does not contain $\tilde{t}$".)

Fisher's exact test calculates the probability of observing the given table or one with a higher joint count:

$$\text{p-value}(CT(\tilde{s},\tilde{t})) = \sum_{k=C(\tilde{s},\tilde{t})}^{\min(C(\tilde{s}),C(\tilde{t}))} p_h(k) \qquad \text{where}$$

$$p_h(k) = \frac{\binom{C(\tilde{s})}{k}\binom{N-C(\tilde{s})}{C(\tilde{t})-k}}{\binom{N}{C(\tilde{t})}}$$

In the following discussion the negative of the (natural) log of the p-value will be computed:

$$\pi(CT) = -\log \text{p-value}(CT)$$

3.2 An example showing how conditioning improves significance calculations

We will show a (not very good) example of this computation that actually occurred in a Giga-FrEn-based phrase table in the experiments discussed below. The phrase pair

(*gouvernement du, of Canada*).

occurs in 44,611 out of the 15,524,575 sentence pairs of the Giga-FrEn corpus. The two by two contingency table is shown in Table 2, and it yields a negative log p-level of 160323.

	oC	oC$'$	
Gd	44611	21786	66397
Gd$'$	145156	15313022	15458178
	189767	15334808	15524575

Table 2: Contingency table for (*gouvernement du, of Canada*) (Here "Gd" will stand for *gouvernement du* and "oC" for *of Canada*)

As this example shows, significance pruning applied to phrase tables can lead to the retention of some obviously poor translations. A more sensitive analysis results if we construct a three by three table (Table 3) from the above by splitting out the counts for *gouvernement du Canada* for the rows and *government of Canada* for the columns. Almost all of the co-occurrence counts for (Gd,oC) is attributable to the (GdC,GoC) phrase pair of which it is a proper sub-phrase-pair.

	GoC	G\oC	oC$'$	
GdC	43800	655	14372	58827
Gd/C	74	82	7414	7570
Gd$'$	10183	134973	15313022	15458178
	54057	135710	15334808	15524575

Table 3: Contingency table for (*gouvernement du Canada, government of Canada*) (GdC,GoC) with excess *gouvernement du* and *of Canada* counts shown (Gd/C≡Counts for Gd with counts for GdC removed and G\oC≡Counts for oC with counts for GoC removed)

To correct the significance level, we can strike out the first row and the first column. This has the effect of conditioning the analysis on those sentence pairs that contain neither a GdC on the French side or a GofC on the English side. The resulting table is Table 4: The significance level for this conditioned table is around 3.6, a value that would not be significant at the 1% level and only slightly significant at the 5% level.

A simpler situation occurs if we only condition on either a longer source phrase or a longer target

	G\oC	oC′	
Gd/C	82	7414	7496
Gd′	134973	15313022	15447995
	135055	15320436	15455491

Table 4: Contingency table for (*gouvernement du, of Canada*) conditioned on the part of the corpus not containing (*gouvernement du Canada, government of Canada*)

phrase and not both. We then have either a two by three table or a three by two table. It will be shown that, even with this limitation, large savings can still be made and translation quality as measured by BLEU (Papineni et al., 2001) actually rises.

Table 5 shows the effect of crossing out row 1 and adding together columns 1 and 2. The significance level is about 41, up from 3.6 but still enough to heavily discount the phrase pair. (Alternatively, the table formed by adding the rows 1 and 2 and deleting column 1 yields a significance level of 442.)

	oC	oC′	
Gd/C	156	7414	7570
Gd′	145156	15313022	15458178
	145312	15320436	15465748

Table 5: Contingency table for (em gouvernement du, of Canada) conditioned on the part of the corpus not containing (*gouvernement du Canada, of Canada*)

In this example, we knew which phrase pair to condition out of the corpus but an automatic process would need to discover this on its own. Because of this we will consider all of the eligible phrase pairs in the phrase table and select the one that has the most effect. The next section will show that we can compute the conditioned contingency table using only the unconditioned tables for the candidate phrase pair and the phrase pair that we are conditioning out.

The *CSignif* column of Table 6 shows a sequence of steps like this where the process is recursively applied to the selected phrase pair. Peak significance levels occur for the sensible phrase pairs (*gouvernement du Canada, government of Canada*) and (*le gouvernement du Canada, the government of Canada*). Pruning in this fashion will improve the use of phrase tables for non-PBSMT applications if this example is a reasonable guide.

3.3 Conditioning of significance testing

We summarize a contingency table (Table 7) for $\tilde{s}$ and $\tilde{t}$ by four numbers ($CT(\tilde{s}, \tilde{t}) = (j, r, c, t)$) and leave blank the positions where the x's appeared, since they will can be calculated by subtraction in all of the contingency tables we consider.

	$\tilde{t}$	
$\tilde{s}$	j	r
	c	t

Table 7: Two by two contingency table for $\tilde{s}$ and $\tilde{t}$ ($CT(\tilde{s}, \tilde{t}) = (j, r, c, t)$)
.

Let $\tilde{t}_i$ be a super-m-gram of $\tilde{t}$, denoted $\tilde{t}_i \sqsupset \tilde{t}$, ($\tilde{t}$ is a subsequence of the sequence of $\tilde{t}_i$). i will range over the rows of the phrasetable having source side $\tilde{s}$. This set of rows will be denoted $\mathfrak{P}(\tilde{s})$. The two by three contingency table is shown in Table 8. Here $j_i = C(\tilde{s}, \tilde{t}_i)$ and $c_i = C(\tilde{t}_i)$.

	$\tilde{t}_i$	$\tilde{t} \mid \tilde{t}'_i$	
$\tilde{s}$	j_i	$j - j_i$	r
	c_i	$c - c_i$	t

Table 8: Two by three contingency table for $\tilde{s}$ and $\tilde{t}$, $\tilde{t}_i$ where $\tilde{t}_i \sqsupset \tilde{t}$

Crossing out the first column of the table gives a contingency table for $\tilde{s}$ and $\tilde{t} \mid \tilde{t}'_i$: $CT(\tilde{s}, \tilde{t} \mid \tilde{t}'_i) = (j - j_i, r - j_i, c - c_i, t - c_i)$ as in Table 9.

	$\tilde{t} \mid \tilde{t}'_i$	
$\tilde{s}$	$j - j_i$	$r - j_i$
	$c - c_i$	$t - c_i$

Table 9: Two by two contingency table for $\tilde{s}$ and $\tilde{t} \mid \tilde{t}'_i$ where $\tilde{t}_i \sqsupset \tilde{t}$

We can do the equivalent thing on the source side: $CT(\tilde{s} \mid \tilde{s}'_k, \tilde{t}) = (j - j_k, r - r_k, c - j_k, t - r_k)$ where $k \in \mathfrak{P}(\tilde{t})$ as in Table 10.

French	English	C(f,r)	C(f)	C(e)	USignif	CSignif
G du	of C	44611	66397	189767	160323	41
G du C	of C	44455	58827	189767	168836	541
G du C	G of C	43800	58827	54057	241191	71717
G du C	the G of C	31761	58827	38759	169516	62570
le G du C	the G of C	20115	27820	38759	110602	89648
le G du C	the G of C is	3749	27820	4340	22252	13615
le G du C est	the G of C is	1407	1874	4340	10722	5709
«le G du C est	the G of C is	649	1874	1099	5243	1223
«le G du C est	" the G of C is	481	522	1099	4580	4887
«le G du C est	, " the G of C is	10	522	12	99	59
que «le G du C est	, " the G of C is	4	4	12	57	43

Table 6: Some Giga-FrEn phrase pairs with counts and significance ($-\log$(p-level)) (G≡*gouvernement* or *government* and C≡*Canada*) The unconditional significance levels are shown in the *USignif* column and conditional significance levels (on the following row) in the *CSignif* column.

	$\tilde{t}$	
$\tilde{s} \mid \tilde{s}'_k$	$j - j_k$	$r - r_k$
	$c - j_k$	$t - r_k$

Table 10: Two by two contingency table for $\tilde{s} \mid \tilde{s}'_k$ and $\tilde{t}$ where $\tilde{s}_k \sqsupset \tilde{s}$

It is useful to consider a degenerate case where a sentence pair from the originating corpus $\mathfrak{C}$, nominally occurring once is considered as super-phrase-pair with the whole target sentence standing in as a super-phrase on the target side. The joint count will be 1 (usually) and the marginal count will 1 (usually) although a more careful analysis could be done. By a similar analysis to above we have in Table 11. By symmetry, the same thing can be done on the source side with identical result. $CT(\tilde{s} \mid \mathfrak{C}', \tilde{t}) = CT(\tilde{s}, \tilde{t} \mid \mathfrak{C}') = (j-1, r-1, c-1, t-1)$. Note that it is possible for $j-1$ to be zero, with our without $r-1$ or $c-1$ being zero. This corresponds to a co-occurrence count of 1. If all three of these values are 1, we have the case of the 1-1-1 contingency table.

	$\tilde{t} \mid \mathfrak{C}'$	
$\tilde{s}$	$j-1$	$r-1$
	$c-1$	$t-1$

Table 11: Two by two contingency table for $\tilde{s}$ and $\tilde{t} \mid \mathfrak{C}'$

3.4 The algorithm for conditional significance pruning

Now with the notation from the last section we can calculate three conditional contingency tables by searching over all possible super-phrases on the target and source side.

$$\begin{aligned} i^* &= \operatorname*{arg\,min}_{\substack{i \in \mathfrak{P}(\tilde{s}) \\ \tilde{t}_i \sqsupset \tilde{t}}} \pi(CT(\tilde{s}, \tilde{t} | \tilde{t}'_i)) \rightarrow CT(\tilde{s}, \tilde{t} | \tilde{t}'_{i^*}) \\ k^* &= \operatorname*{arg\,min}_{\substack{k \in \mathfrak{P}(\tilde{t}) \\ \tilde{s}_k \sqsupset \tilde{s}}} \pi(CT(\tilde{s} | \tilde{s}'_k, \tilde{t}) \rightarrow CT(\tilde{s} | \tilde{s}'_{k^*}, \tilde{t}) \\ & \qquad\qquad\qquad\qquad\qquad \rightarrow CT(\tilde{s}, \tilde{t} | \mathfrak{C}') \end{aligned}$$

To the right of the arrows, the three lines in this formula show three candidate contingency tables formed by conditioning on a containing phrase pair where the source phrase is held constant, where the target phrase is held constant, and conditioning as in Table 11. We will then choose one of these three, the one that is least significant, that has the smallest value of $\pi(\cdot)$.

In the next section, three approaches will be evaluated: (1) the original significance pruning, (2) conditional significance pruning using only the first two cases if defined, falling back on unconditional pruning if there is no qualified super-phrase on either the source or target side, and (3) full conditional significance pruning, as just described. Note that case (3) will automatically discard any phrase pairs with a co-occurrence count of 1, including all of the 1-1-1's.

4 The Experiment

4.1 The baseline statistical machine translation system

The Moses baseline system for WMT11 (Workshop on Statistical Machine Translation, 2011) along with corpora from the WMT11 evaluation were employed. Changes were only made to ensure that scripts would run correctly, and the use of training and dev sets mirrored those of the WMT11 evaluation. Thus the results from this study conform exactly to the WMT11 conditions except for the timing and much larger time for training and tuning.

A snapshot of Moses (Koehn et al., 2007) was taken on August 3, 2011. This was combined with then current versions of SRILM (Stolcke, 2002) (v. 1.5.12), GIZA++ (Och and Ney, 2000) (v1.0.5), and the mteval script used for WMT11 (mteval-v11b.pl).

The directions provided on the WMT11 website for baseline system 1 were followed, to produce 5 versions of a French to English Baseline system. The tokenize.perl script supplied with for the WMT11 exercise was used in all cases. Adherence to the directions was strict to ensure reproducibility and to study the effects of pruning on an un-tuned strong baseline.

Optimal weights for these features were trained using n-Best MERT as provided with Moses. The dev set used for training was the WMT09 test set. The WMT10 test set was used for a dev test to choose the best configuration and final evaluation is based on the WMT11 eval set. This is consistent with competition in the WMT11 shared task (although BLEU is only an auxiliary measure for evaluation and human evaluation is the official measure).

The baseline configuration does not use Kneser-Ney smoothing on the phrase table but does use Kneser-Ney smoothing on the 5-gram language model. Since this evaluation is of phrase tables, it is important to ensure that the choice of language model does not skew the results. In each case, the language model was computed on the un-duplicated monolingual English corpus composed of all of the news.

This is a large in-domain monolingual corpus. No language models derived from the other bilingual corpora were used. Because the results of this looked reasonable and this study is about phrase tables, this decision was taken after experimentation not reported here. A test was done including, in addition, the corresponding English side of the parallel corpora, but this did not improve BLEU. The LDC supplied corpora were not used to stay within the WMT11 supplied materials.

The 5 phrase tables were produced using Giza++ and the diag-end-final algorithm in Moses. Because there are 4 natural sets of bilingual data for French to English, there were four separate phrase tables produced: (1) News-commentary (small in-domain), (2) Europarl (larger not in-domain but closer than the other larger corpora), (3) UN, and (4) Giga-FrEn (the web-harvested huge bilingual corpus). There was a fifth phrase table produced by pooling the four corpora into one large corpus and producing a phrase table.

In order for the Moses phrase table scripts to handle such large corpora, an initial filtering step in applied where long tokens and long phrases are censored as well as phrase pairs with an extreme length ratio.

A final note: although only one replicate per point is shown in the results, every data point is the result of a separate MERT training run and so they can be interpreted as independent trials on slightly different conditions. In all pruned conditions the starting weights for MERT were the optimized weights for the 100% case.

5 Results

5.1 Confirmation of unconditional pruning behaviour

The first question to investigate is the issue of whether the original significance pruning does not penalize BLEU in the case of this data set. Some authors (Yang and Zheng, 2009) have noticed that it is sometimes the case that some of the 1-1-1's need to be included in order to achieve the optimum balance between translation quality and amount of pruning. The baseline is also stronger since it includes a number of enhancements made since 2007.

Figure 1 shows the effect of retained percentage of the phrase table on BLEU for each of the five corpora. It is quite clear that for News-Commentary there is a cost in pruning that is serious and would

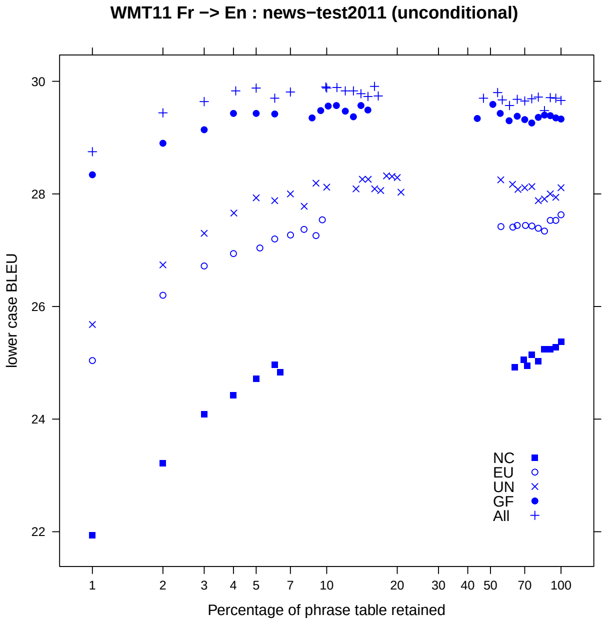

Figure 1: WMT11 news-test2011 (unconditional pruning)

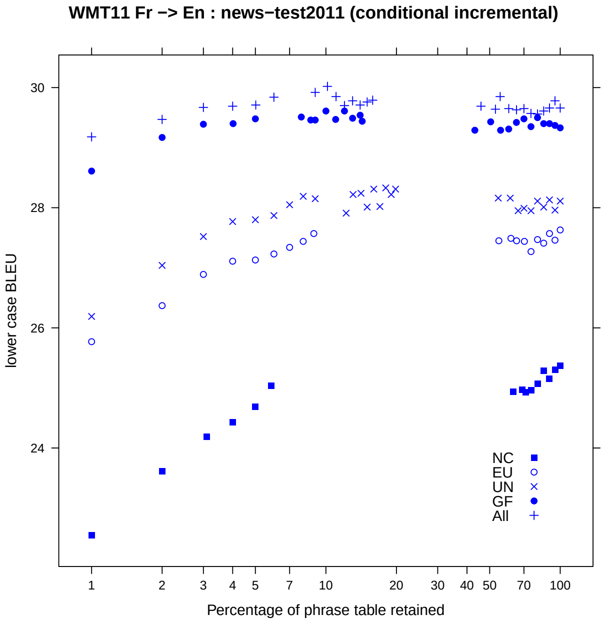

Figure 2: WMT11 news-test2011 (unconditional pruning updated with conditional levels when defined)

cause hesitation to using significance pruning. For Europarl, there is a tiny cost that might be considered serious or not depending on the application. For the larger corpora though, there is quite clearly a benefit from the standard significance pruning, discarding all of the 1-1-1's and those phrase pairs less significant.

Note that the large gap in each set of points corresponds exactly to the 1-1-1's.

5.2 Behaviour of conditional significance pruning

The two conditional significance pruning schemes are applied to the same corpora. The first scheme with the significance levels updated with conditional significance levels where they exist is shown in Figure 2.

The graphs look very similar to the unconditional graphs although they have slightly higher BLEU scores. This is not easy to see though and will be studied below. In particular the large 1-1-1 gap is still visible and there is a loss of BLEU performance if 1-1-1 and below phrase pairs are discarded for NC and EU.

The total conditional significance pruning scheme discards all of the 1-1-1's and there are two gaps. The right-most gap corresponds to all phrase pairs with a significance p-value of 100%. This includes all of the 1-1-1's as well as some degenerate conditional contingency tables. The other gap corresponds to all phrase pairs that are anti-associated or have a significance p-value of 50% or above.

Figure 3 shows that the peak translation quality occurs just to the left of the two gaps in all cases except News-Commentary where there still is a small benefit from including the 1-1-1's. Notice the improvement in BLEU by retaining about 20% of the conditionally pruned table compared to about 10% of the unconditionally pruned table.

5.3 Improvement of conditional over unconditional pruning

The runs were set up to match the percentages of corpus retained to allow for a paired t-test. To visualize the effect the value of BLEU from unconditional pruning is subtracted from the BLEU from the conditional pruning under the two schemes. The results are shown in Figure 4 and Figure 5 for pruning to 20% or less.

It is quite clear from the graphs that both of the schemes show an improvement and that it is significant. The average difference is 0.09 BLEU for

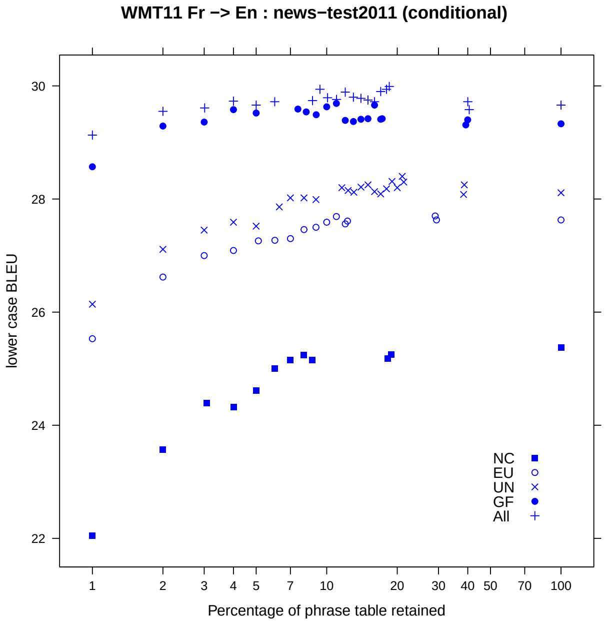

Figure 3: WMT11 news-test2011 (conditional pruning)

scheme 1 and 0.07 for scheme 2. These are both significance at a 1% level. Nonetheless, it is clear from the graphs that the actual improvement depends on the prune percentage.

It appears from this analysis that scheme 1 is better than scheme 2 but earlier graphs show that this is not the case for the larger corpora at the 50% threshold.

5.4 How would this system have done for WMT11?

The evaluation for WMT11 was based on a human evaluation. However, if we adopt a proper protocol of deciding scenarios based on news-test-2010 and evaluating on news-test-2011, we can compare the results against the scores published on the web site. The best lower case BLEU on the eval set was posted as 30.5. The first interesting observation is that the baseline 1 system can achieve 29.66 BLEU by combining all of the parallel corpora. This also established that this is a strong baseline.

Table 12 shows the results if no pruning is done. Eval BLEU scores are in bold if they are within 0.1 BLEU of the best achieved for this corpus. For the smaller corpora NC and EU, pruning provides no benefit and for NC can cause loss in BLEU. This agrees with the results from a number of papers that agree that some of the 1-1-1's need to be included. In the case of EU it is less clear as pruning can almost reach the un-pruned case.

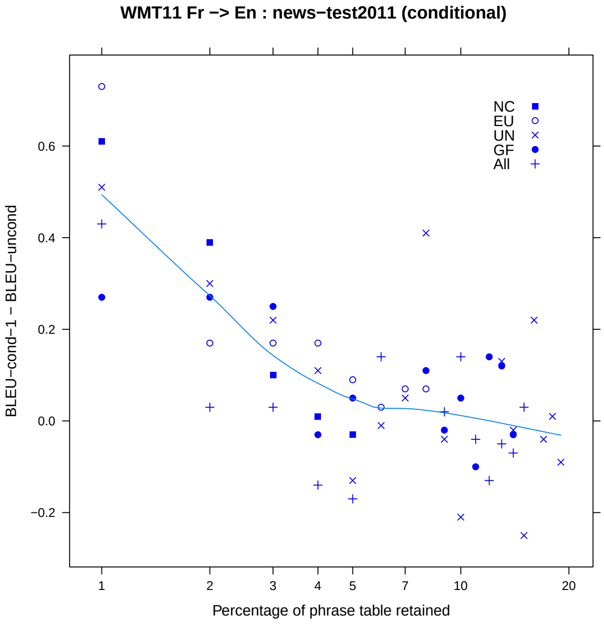

Figure 4: WMT11 news-test2011 BLEU-cond-1 - BLEU-uncond

Corpus	Test BLEU	Phrase Table %	Eval BLEU
NC	23.49	100.000	**25.37**
EU	27.30	100.000	**27.63**
UN	26.14	100.000	28.11
GF	28.39	100.000	29.33
All	29.03	100.000	29.66

Table 12: Strategy 0 : No Pruning

A common strategy choosing the best configuration is based on the test BLEU. One notable observation is that, except for GF, the best test BLEU is not achieved without pruning. This is different behaviour from the eval BLEU as can be seen from the results earlier in this section. The results are shown in Table 13.

However, pruning the phrase table to less than 20% can improve the ability of the model to generalize to the eval set. Table 14 shows the results using a strategy of choosing the pruned model with the best test BLEU.

The NC corpus loses about 0.2 BLEU and UN, GF, and All gains about 0.2 BLEU. EU is about the

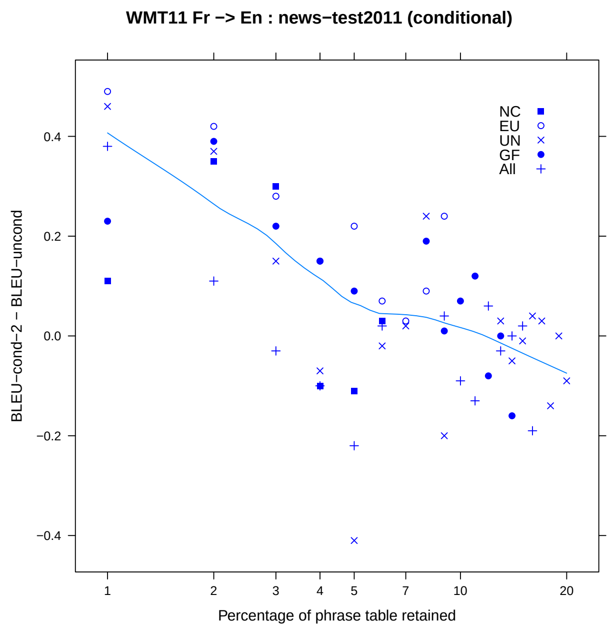

Figure 5: WMT11 news-test2011 BLEU-cond-2 - BLEU-uncond

Corpus & Scheme	Test BLEU	Phrase Table %	Eval BLEU
NC1	23.49	95.000	**25.28**
NC2	23.53	95.005	**25.30**
NC3	23.49	100.000	**25.37**
EU1	27.31	80.005	27.39
EU2	27.43	75.060	27.27
EU3	27.47	29.417	**27.63**
UN1	26.14	100.000	28.11
UN2	26.14	100.000	28.11
UN3	26.14	100.000	28.11
GF1	28.79	14.000	**29.57**
GF2	28.71	14.000	**29.54**
GF3	28.71	8.177	**29.54**
All1	29.08	95.000	29.70
All2	29.05	80.000	29.56
All3	29.03	100.000	29.66

Table 13: Strategy 1 : Best Test BLEU

same pruned as not.

Oracle best BLEU **30.02** happens for All2 with 10% and test BLEU of 28.93. There is no obvious scenario under which this configuration would be chosen. However, someone believing fully in the merits of method 3 and the superiority of using all of the corpora pooled together might have chosen to select all positively associated phrase pairs (significance p-level = 50%). In this case 18.492% of the phrase table would be selected (157,503,821 phrase pairs). The BLEU on the test set in this case is 28.91 and on the eval set it is **29.99**.

5.5 The running cost

The run of All un-pruned required 38GB of RAM for 3 or 4 hours on a R710 to translate the test or eval set. The three pruned runs required about 14GB of RAM for 1 to 1.5 hours on a R710.

This is the other reason for pruning. It allows much more translation and experimentation to be achieved with the same resources.

6 Conclusions

From the results it is clear, at least for this huge corpus, that significance pruning is effective at reducing the size of the phrase table to less than 20% of its size and deliver a small consistent improvement to BLEU. The modification of conditioning can raise this to up to a 0.33 BLEU improvement. This is a strong baseline because it would have placed in the top few of the WMT11 systems using only resources available at that time.

For the Europarl corpus, the results are similar to those in the significance pruning paper (Johnson et al., 2007).

Future work could include studying the effect of including some of the 1-1-1's for huge corpora based on some of these techniques, but the above results do not suggest that this will yield large benefits.

A more promising avenue of research involves combining the ideas of significance testing more intimately into the phrase generation process. If the corpus does not support some of the phrases strongly, maybe it also does not support the word alignments that lead to these phrases. This insight might lead to improvements in word alignments.

Another interesting direction is the use of high quality phrase equivalents for other purposes. As the quality improves, techniques that depend on translation equivalents may also improve.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. 1993. The Mathemat-

Corpus & Scheme	Test BLEU	Phrase Table %	Eval BLEU
NC1	22.85	6.324	24.83
NC2	22.91	5.844	25.04
NC3	23.31	8.696	25.15
EU1	27.19	9.577	**27.54**
EU2	27.27	8.881	**27.57**
EU3	27.42	12.266	**27.61**
UN1	25.94	18.002	**28.32**
UN2	26.03	19.832	**28.31**
UN3	26.11	19.000	**28.31**
GF1	28.79	14.000	**29.57**
GF2	28.71	14.000	**29.54**
GF3	28.71	8.177	**29.54**
All1	29.04	16.000	**29.91**
All2	28.99	15.834	29.79
All3	29.01	14.000	29.78

Table 14: Strategy 2 : Best Pruned Test BLEU

ics of Statistical Machine Translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312, June.

Yu Chen, Andreas Eisele and Martin Kay. 2009. Intersecting multilingual data for faster and better statistical translations. In *Proceedings of the 2009 NAACL-HLT*, pages 128–136, Boulder, Colorado, USA.

Nan Duan, Mu Li, Ming Zhou and Lei Cui. 2011. Improving Phrase Extraction via MBR Phrase Scoring and Pruning. In *Proceedings of MT Summit 2011*, pages 189–197.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Fei Huang and Bing Xiang. 2010. Feature-Rich Discriminative Phrase Rescoring for SMT. In *Proceedings of the 2010 Coling*, pages 492–500, Beijing, China.

J Howard Johnson, Joel Martin, George Foster and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. In *Proceedings of Conference on EMNLP 2007*, pages 967–975, Prague, Czech Republic.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1995*, pages 181–184, Detroit, Michigan. IEEE.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Eduard Hovy, editor, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133, Edmonton, Alberta, Canada, May. NAACL.

Philipp Koehn 2003. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Unpublished draft. see `http://www.iccs.inf.ed.ac.uk/˜pkoehn/publications/europarl.pdf`

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexander Constantin and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007*, pages 177–180, Prague, Czech Republic.

Robert C. Moore. 2004. On Log-Likelihood-Ratios and the Significance of Rare Events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440-447, Hongkong, China.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of Machine Translation. Technical Report RC22176, IBM, September.

Germán Sanchis-Trilles, Daniel Ortiz-Martiínez, Jesús González-Rubio, Jorge González and Francisco Casacuberta. 2011. Bilingual segmentation for phrasetable pruning in Statistical Machine Translation. In *Proceedings of EAMT 2011*. pages 257–264, Leuven, Belgium.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, Denver, Colorado, September.

Nadi Tomeh, Nicola Cancedda and Marc Dymetman. 2009. Complexity-Based Phrase-Table Filtering for Statistical Machine Translation. In *MT Summit 2009*, Ottawa, Canada.

WMT 11. 2011. NAACL Workshop on Statistical Machine Translation. see `http://www.statmt.org/wmt11/translation-task.html`

Mei Yang and Jing Zheng. 2009. Toward Smaller, Faster, and Better Hierarchical Phrase-based SMT. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 237–240, Singapore.

Richard Zens, Daisy Stanton and Peng Xu. 2012. A Systematic Comparison of Phrase Table Pruning Techniqes. Accepted for EMNLP 2012. Jeju, Korea.