Complexity-Based Phrase-Table Filtering for Statistical Machine Translation

Nadi Tomeh[(a)] and **Nicola Cancedda**[(b)] and **Marc Dymetman**[(b)]
[(a)]LIMSI-CNRS and Univ. Paris-Sud XI, Orsay
[(b)]Xerox Research Centre Europe, Grenoble
nadi.tomeh@limsi.fr, {cancedda,dymetman}@xrce.xerox.com

Abstract

We describe an approach for filtering phrase tables in a Statistical Machine Translation system, which relies on a statistical independence measure called *Noise*, first introduced in (Moore, 2004). While previous work by (Johnson et al., 2007) also addressed the question of phrase table filtering, it relied on a simpler independence measure, the *p-value*, which is theoretically less satisfying than the Noise in this context. In this paper, we use Noise as the filtering criterion, and show that when we partition the bi-phrase tables in several sub-classes according to their complexity, using Noise leads to improvements in BLEU score that are unreachable using *p-value*, while allowing a similar amount of pruning of the phrase tables.

1 Motivation

Currently, the most widely used Statistical Machine Translation systems are so-called "phrase-based" systems; they are based on tables of "bi-phrases", that is, pairs of the form <source-phrase, target-phrase>, which are learned automatically from bilingual corpora (see (Lopez, 2008) for an overview). While most such systems use contiguous bi-phrases, we are using one, MATRAX, (Simard et al., 2005) that employs non-contiguous bi-phrases such as <*ne ... plus, does not ... anymore*>, which are better able to generalize over certain linguistic patterns. However such bi-phrases may also lead to larger, more combinatorial, tables: while the potential number of contiguous phrases in a sentence grows quadratically in the length of the sentence, that of non-contiguous phrases grows exponentially. It is especially important to control the proliferation of bi-phrases in this situation, in order to reduce the size of the bi-phrase table and also to improve translation performance by removing "spurious" bi-phrases which do not have good predictive linguistic value.

Working in their case with a system based on contiguous bi-phrases, (Johnson et al., 2007) were able to prune many bi-phrases out of the table without negatively impacting the end results, while at the same time improving the decoding speed. Their approach was to assess the strength of the statistical dependence between the source and the target of the bi-phrase, using a "*p-value*" measure based on a standard independence test, and to prune from the table those bi-phrases for which this strength was below a certain threshold.

We innovate on several important aspects relative to this prior art.[1] *First*, rather than filtering on the basis of *p-value*, we filter based on a different measure of statistical dependence, namely the so-called *"Noise"* introduced in (Moore, 2004). This measure is in principle superior to the simpler *p-value* for the situation at hand: while the *p-value* estimates the statistical dependence between the source-phrase and the target-phrase, based on the corpus statistics associated with *this individual bi-phrase*, the Noise

[1]There are some other approaches to pruning bi-phrases, such as the method described in (Eck et al., 2007), where pruning is determined by bi-phrase usage statistics during decoding. Here our focus is on techniques based on statistical significance tests.

takes into account this pair *in the context of all other bi-phrases*, which is in theory better statistically motivated in the context of large bi-phrase libraries, as is the case in SMT. *Second*, we show that, when we distinguish different classes of bi-phrases according to their complexity (roughly, their size), then thresholding on Noise makes different predictions than when thresholding on *p-value* and produces *better SMT results* while permitting a similar level of pruning. *Third*, we introduce a simple *simulation-based* approach for computing the Noise: we directly compare the statistics observed in the corpus to statistics obtained by simulating a virtual "randomized" corpus that has similar statistics to the observed corpus, but in which the translation correlations between source and target sentences have been neutralized through permutation; this is simpler than previously described techniques and is also more flexible, since it can be used to compute different variants of the "null hypothesis" used for assessing the significance of the bi-phrases in the table.

2 Background: Computing Association Scores Through Fisher's Exact Test

Before moving to our approach, we start by describing the background approach where one computes association scores between source- and target-phrases based on the *p-value* associated to a certain statistical independence test, the so-called "Fisher Exact Test" (Agresti, 1992). This test computes the probability ("*p-value*") that a certain joint event (A,B) appears under a so-called "null hypothesis" that correspond to the situation where A and B are statistically independent. In the case where the joint event under consideration is the joint occurrence of a source phrase S and a target phrase T in respectively the source and target side of the same bi-sentence, the meaning of the *p-value* can be explained by considering the following protocol: Given a S, given a T, given that there are N bi-sentences in the corpus, given that S appears in $C(S)$ bi-sentences and T in $C(T)$ bi-sentences, given the (null) hypothesis that the $C(S)$ occurrences of S are placed independently at random in the corpus and similarly for the occurrences of T, *then,* what is the probability that the joint occurrences of (S,T) will appear $C(S,T)$ times or more?

This probability is called the *p-value* associated with the *contingency table*:

$C(S,T)$	$C(S)-C(S,T)$
$C(T)-C(S,T)$	$N-C(S)-C(T)+C(S,T)$

Fisher's Exact Test computes this *p-value* exactly, using the following formulas (where p_{hg} denotes the hypergeometric distribution):

$$p_{hg}(k) = \frac{\binom{C(S)}{k}\binom{N-C(S)}{C(T)-k}}{\binom{N}{C(T)}}$$

$$\textit{p-value}(C(S,T)) = \sum_{k=C(S,T)}^{\min(C(S),C(T))} p_{hg}(k)$$

The smaller this *p-value*, the more significant the "dependence" between S and T is considered to be. Indeed, when the *p-value* is close to 0, then the probability of finding as many joint occurrences as $C(S,T)$, given that the marginals are $C(S)$ and $C(T)$, and given that the occurrences of S and of T are placed at random among the N bi-sentences, is close to 0. We also define the association-score relative to a contingency table:

$$association\_score \equiv -log(\textit{p-value}),$$

which varies from 0 to ∞, with high numbers indicating a strong statistical dependence.

Of course, Fisher's Exact Test is not the only statistical test of independence around, a better known one being the χ^2 test; however, while more computationally costly, the exact test is more accurate than the χ^2 when the counts in the first three contingency table cells are small, which is typically the case in application to phrase tables. Note that, at points in the experiments where we have to compute the *p-value*, we rely on the efficient R implementation of Fisher's exact test (R-Manual, 2009).

3 A Problem with Association Scores

In a paper devoted to rare-event associations in a bilingual corpus (Moore, 2004), Moore noted a problem about standard statistical significance scores such as the *p-value* defined above. To explain the problem, we keep the same notation as before, although (Moore, 2004) concentrates on associations between words rather than phrases. Here is the problem: a high nominal association score

between S and T does not always indicate dependence. Consider the following example. Suppose that the corpus contains $N = 500,000$ bi-sentences; call a word S (resp. T) a singleton if it is represented once in the source (resp. target) side of the corpus; suppose that we observe $17,379$ source singletons and $22,512$ target singletons; suppose also that we observe $19,312$ s-s (singleton-singleton) pairs. Note that each such pair has the same contingency table, and hence the same high association score $log(p\text{-}value) = -log(1/500,000)$. However, if we placed these singletons independently at random among the $500,000$ bi-sentences (on the source and target sides resp., our null hypothesis here), we would expect to observe around: 782.5 ($= 17,379 \times 22,512/500,000$). This means that if we had observed around 782 s-s pairs (rather than $19,312$), we should have absolutely no confidence that any such pair (S,T) is actually indicative of a statistical dependence between S and T, despite the high association score of $-log(1/500,000)$.

In other words, one should be careful when interpreting such association scores. For any *given* singleton-singleton pair it is indeed true that the probability of it occurring by chance if the two singletons are actually statistically independent is $1/500,000$. But we would be wrong to conclude from this that the fraction of the global population of singleton-singleton pairs that were observed (namely $19,312$) which is due to chance is only $1/500,000$. Indeed, this would be the case if s-s pairs were statistically independent from one another, but clearly they are not, and the fraction of unreliable s-s associations is in general much larger.

In order to remedy this problem, Moore introduces the notion of *Noise*, which is the ratio between the expected number of s-s pairs under the independence assumption and the actually observed number of s-s pairs. In our example, we have: $Noise = 782.5/19,312 \simeq 4\%$. If we had observed around 782 s-s pairs, the Noise would be close to 100%. However, given that we have actually observed $19,312$ such pairs, we may say that there is about 0.04 "probability" that a given s-s pair is due to chance; note that using the raw association score to estimate this probability would give us the *p-value* $1/500,000 = 0.000002$, that is, a much too optimistic estimate of the dependence.

We take the Noise as our new indicator of deviation from independence for a s-s pair (S,T): *low Noise indicates strong dependence*.

Beyond singletons: general definition of Noise Consider a corpus $\boldsymbol{C}$ of N bi-sentences; also consider a bag of source words $\boldsymbol{S}$, and a bag of target words $\boldsymbol{T}$, where the elements of $\boldsymbol{S}$ and $\boldsymbol{T}$ can be represented several times in the corpus (so we are not anymore limited to singleton words). If β is an *association-score* level, define: $observed(\beta)$= the number of different pairs (S,T) (with S in $\boldsymbol{S}$, T in $\boldsymbol{T}$) such that: *association-score*$(C_{obs}(S,T), C(S), C(T), N) > \beta$ where C_{obs} is the observed count of (S,T) in $\boldsymbol{C}$, and: $expected(\beta)$= the number of different pairs (S,T) (with S in $\boldsymbol{S}$, T in $\boldsymbol{T}$) such that: *association-score*$(C_{exp}(S,T), C(S), C(T), N) > \beta$ where C_{exp} is the expected count of (S,T), assuming an independent "generative" model where each S in $\boldsymbol{S}$ (resp. each T in $\boldsymbol{T}$) is placed at random in $[1,\ldots,N]$. Then (Moore, 2004) defines:

$$Noise(\beta) = expected(\beta)/observed(\beta)$$

The smaller $Noise(\beta)$, the more "signal" there is in the corpus about dependencies between the words of S and of T.

Note that, whenever the *p-value* of a statistical significance test like Fisher's exact test is used as an association score, the previous discussion is an instantiation to the case of bilingual associations of a general statistical procedure called *multiple hypotheses testing*, or *multiple comparisons*. Indeed, Moore's "Noise" is related to "the proportion of false discoveries among the discoveries" introduced in (Soric, 1989) (as reported in (Benjamini and Hochberg, 1995).[2]

4 Noise-Based Pruning of Bi-phrases

[2]The relation is however not one of equivalence. In the case of Noise, while the number of "discoveries" is modeled by $observed(\beta)$, the number of false discoveries among them is modeled as $expected(\beta)$, that is the number of discoveries that one would expect to make at threshold β if *all* the bi-phrases were actually independent, whereas Soric's "proportion of false discoveries among the discoveries" would actually degenerate to a value close to 1 in this case (because all discoveries on the observed corpus would be false discoveries), that is, would be useless.

Bi-phrase pruning in (Johnson et al., 2007) (Johnson et al., 2007) filter phrase tables directly based on the p-value of Fisher's exact tests, not on Noise: a bi-phrase is pruned if its p-value is above a certain threshold γ, tuned on a validation set to optimize the balance between translation quality and phrase-table size. This is justified in their setup, since Noise behaves monotonically in the p-value, and thresholding the former is equivalent to thresholding the latter.

Noise-based bi-phrase pruning Our approach is based on the following remark: there is a certain fallacy in directly transposing Moore's Noise computations from words to phrases. This is because phrases may vary considerably in complexity, from one single word to several words, and additionally, when considering non-contiguous phrases, from phrases with no gaps to phrases with multiple gaps. In such circumstances, contrast the following two situations: (i) two words S and T appear in the same bi-sentence B, respectively on its source and on its target, and otherwise these words appear nowhere else; (ii) two complex phrases S' and T' appear in the same sentence, and otherwise these phrases appear nowhere else.

First note that we have the same contingency table for S, T as we have for S', T', namely:

1	0
0	$N-1$

and therefore we also have the same association score in the two situations.

However, these situations are quite different: in the second case, it is quite common for complex phrases such as S' and T' to appear only once overall in the corpus (because, the more complex a phrase, the rarer it is), and so we should not be at all surprised that S' and T' appear only in B — for instance such observations would naturally occur with a high frequency even if the source and sentence side of the corpus were permuted randomly; in the first case, by contrast, words often tend to occur several times in the corpus, and therefore the fact that S and T only occur in B is more interesting: such things would not occur so frequently in a random permutation of the corpus.

Thus, while the association scores for (S, T) and (S', T') are exactly the same, we would be mistaken in believing that this fact gives us the same evidence as to their true statistical dependence. However, the approach in (Johnson et al., 2007) would either prune both of them from the table, or none of them.

Complexity classes Our approach is the following. Rather than computing $Noise(\beta)$ uniformly for all possible bi-phrases, we partition the bi-phrases into several *complexity classes*, and for each such class we compute $Noise(\beta)$. In more detail, here is how we proceed:

(1) We choose a certain number of threshold levels for the association score β; this is done for computational convenience in order to "discretize" the computation, in principle we can use as fine a grid of thresholds as desired.

(2) We build a table of (non-contiguous) bi-phrases, based on the bilingual corpus, using the procedure described in (Simard et al., 2005).

(3) We partition the global bi-phrase table ***U*** into $N = 4$ subtables ***L1***, ***L2***, ***L3***, ***L4*** of increasing complexities: bi-phrases composed of 1 "cept", 2 cepts, 3 cepts or 4 cepts, where cepts are "elementary" bi-phrases.[3]

(4) For each subtable we project its bi-phrases into their source phrases and target phrases, obtaining a bag of source phrases S and a bag of target phrases T. For instance, the bi-phrase $<$*ne ⋄ plus; not ⋄ ⋄ ⋄ anymore*$>$ is projected into the source phrase $<$*ne ⋄ plus*$>$ and into the target phrase $<$*not ⋄ ⋄ ⋄ anymore*$>$.

(5) We compute, for each threshold β, the number of bi-phrases formed from elements of ***S*** and ***T*** that have an association score larger or equal to β. These counts are based on our original bilingual corpus, and are called the *observed counts*.

(6) We now perform exactly the same computation as in the previous step, but this time based on a "virtual" corpus, which is obtained from the original corpus through a randomized simulation that "neutralizes" the translation dependencies present in the original corpus (this simulation process is described in more detail below). The counts obtained are called the *expected counts* under the independence hypothesis.

(7) We compute $Noise(\beta)$ as the ratio between

[3] In MATRAX, a "cept" is an elementary bi-phrase from which more complex bi-phrases can be built. For instance, the two cepts $<$*drank; a bu*$>$ and $<$*wine; du vin*$>$ might be combined into the "two-cept bi-phrase" $<$*drank wine; a bu du vin*$>$.

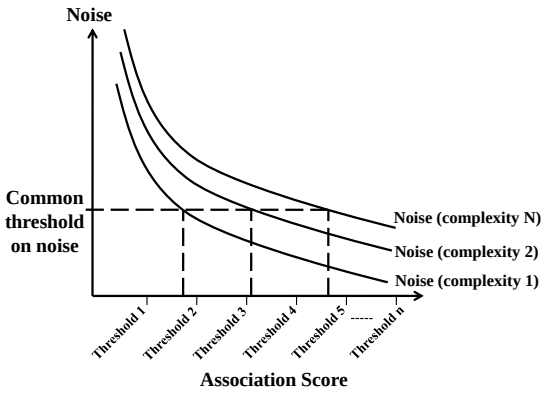

Figure 1: When computing Noise curves separately on subtables of different complexities, a single Noise threshold corresponds to different thresholds for the different subtables.

expected counts and observed counts.

(8) As a result, we obtain four curves relating the Noise to the association score, one for each complexity class (see Figure 1).

We see that, now, a common threshold on the Noise across the several complexity classes corresponds to a different association score for each complexity class. Rather than pruning the bi-phrases based on a given association score, we will now prune them based on a common Noise level, and this constitutes an essential difference with (Johnson et al., 2007), with Noise now playing an effective role.

Simulating a corpus under some independence hypothesis In (Moore, 2004), the author shows how to compute, for words S and T that have marginal counts $C(S)$ and $C(T)$, the expected number of times S and T would be found in the same bi-sentence, assuming a certain independence hypothesis which corresponds to the following generative model: for each of the $C(S)$ occurrences of S, put the occurrence at random among the N bi-sentences of the corpus, and similarly for each of the $C(T)$ occurrences of T. Under this simple model, it is possible to analytically compute the expected number of times S and T are found in the same bi-sentence, namely $C(S) * C(T)/N$.

Although we could apply the same kind of computation to the situation with phrases, rather than words, we prefer to *simulate* a virtual corpus, and estimate the bi-phrase "expected counts" by simply observing bi-phrase counts in this virtual corpus.[4] This gives us more freedom as to the exact nature of the "null hypothesis" providing the baseline for the independence test between source and target phrases. For instance we might want to keep invariant the *word-level* statistics in the corpus (that is, preserve the counts of each source or target word) while randomly permuting the source words across the source sentences and similarly for the target words, and then we might want to compare the bi-phrase counts obtained on this simulated corpus ("expected counts") with the bi-phrase counts obtained in the actual corpus.

In our experiments, the simulation that we use is the following one: we keep the set of source sentences and the set of target sentences identical to those of the original corpus, but we permute the positions of the target sentences randomly relative to those of the source sentences, thereby destroying the translation relation between the source and the target sentences. The procedure has the advantage that the source phrases ***S*** and the target phrases ***T*** that were listed in the bi-phrase table for the original corpus keep the marginal counts that they had there. The observed count of (S, T) in the simulated corpus is just the number of times that that pair appears in some bi-sentence of the simulated corpus. In addition, this procedure preserves the statistical dependencies between phrases in the same language. An extreme example of such dependencies is given by those cases when a phrase S (and similarly for T) is a complex phrase: a sentence containing S will also necessarily contain all subphrases S' of S, and it would not be correct to assume that occurrences of S and S' can be placed independently; in the case of Moore's approach, this phenomenon is less perceptible, because he discusses single words rather than phrases, and randomly placing words in different sentences ignoring statistical dependencies be-

[4]The wording "expected counts" is then a slight misnomer in this situation, since we typically look at only one simulated corpus, rather than at several instances of such corpora, produced from the actual corpus through the same generative process, over which expected counts could be estimated through averaging. In practice, with a large enough original corpus, the difference between using such an averaging process and only using one simulation is negligible.

tween words of the same language is a less severe simplification. In the case of phrases as opposed to words, if we placed the phrases randomly among the corpus sentences independently of each other, we would significantly modify the statistical properties of the source (resp. target) sentences. Using the current approach, we do not change these statistical properties, but concentrate on variations due only to the randomization of target sentences *relative* to source sentences.[5]

5 Experiments

First experiment: Pruning based on association-score Our first, baseline experiment, uses raw association scores for pruning the bi-phrases. It is similar to what (Johnson et al., 2007) did, but in our case with (i) non-contiguous bi-phrases, (ii) a relatively small corpus with high lexical homogeneity (French-English Technical Documents, 52,322 parallel sentences, 632,753 French tokens, 570,340 English tokens).

Table 1 shows some variants of the bi-phrase table constructed on this corpus. On each variant of the table, the same procedure was applied: for each of a few predetermined threshold levels on the association score, a bi-phrase was pruned from the table according to whether its association score (computed from the contingency table of this bi-phrase in the corpus) was higher or lower than the threshold. Figure 2 shows the results for the different table variants. The vertical axis represents the number of bi-phrases that were kept in the table, while the horizontal axis represents the association score threshold. The special level called α represents the association score obtained for a contingency table of the form $(1, 0, 0, N - 1)$, this level having the property that any bi-phrase having an association score strictly larger than α necessarily appears at least twice in the corpus. We then conducted experiments in order to assess the impact on translation performance of the pruning threshold. In Figure 3

Bi-phrase table	Number of bi-phrases
U^1	77,370
$U^2 - g0$	125,379
$U^2 - g4$	340,034
$U^3 - g4$	534,107
$U^4 - g4$	569,471

Table 1: Some variants of the bi-phrase table in MATRAX. U^1 is made of bi-phrases containing only one cept, with an arbitrary number of gaps, $U^2 - g0$ bi-phrases containing up to 2 cepts, but with no gaps, $U^2 - g4$ bi-phrases containing up to 2 cepts with at most 4 gaps, $U^3 - g4$ similarly with 3 cepts, and $U^4 - g4$ with 4 cepts.

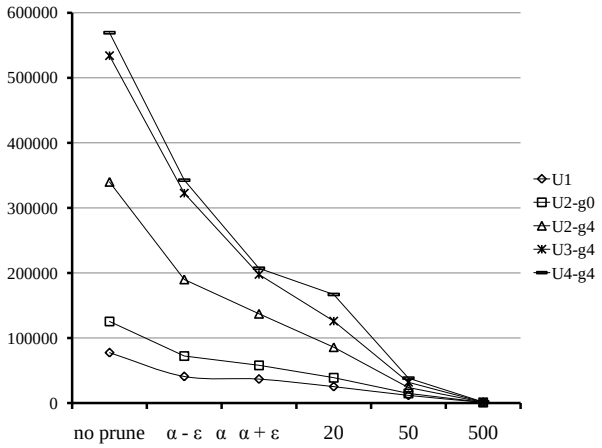

Figure 2: Effect of the association score on the level of pruning for different bi-phrase tables. The effect of pruning at a given threshold is different across tables, with a steeper effect for the more complex tables.

[5] One reviewer has been skeptical about whether this dependency between source phrases S, S' or target phrases T, T' had any impact on the value of $expected(\beta)$, and in consequence, whether that value (or more precisely, the true expectation of that value rather than the estimate based on a single simulation) would be different from the value computed through the analytical method of Moore, taking phrases rather than words as the elementary units. Although it is easy to show that the *distributions* of contingency tables can be very different under the two models, we now tend to share the reviewer's opinion that the *expectations* over such distributions that are relevant for the computation of $expected(\beta)$ may be identical or at least similar (due to minor technical differences). This would require a careful check, but, even if that were the case, it is worth noting how *flexible* our simulation approach is, and there are several variants that we would like to explore in future work. To give an extreme example, we could first randomly permute the words across source sentences and target sentences, then decide to completely rebuild the bi-phrase table starting from *that* corpus, and finally use the statistics obtained on *those* bi-phrases for computing the Noise of the original corpus relative to the simulated corpus. There is no way this kind of power could be attained by using simple analytical formulas for the expected counts of the kind Moore uses: the construction of a bi-phrase table from a corpus is not amenable to such closed-form solutions.

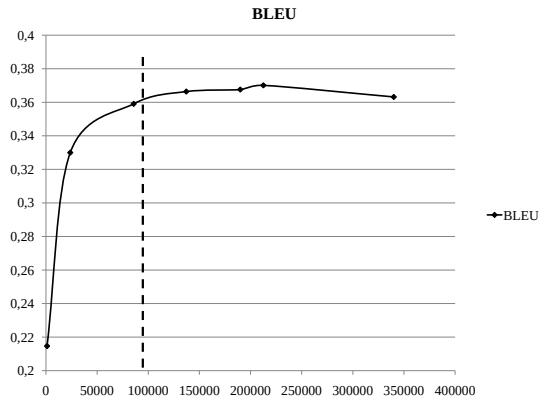

Figure 3: Translation performance relative to the level of pruning.

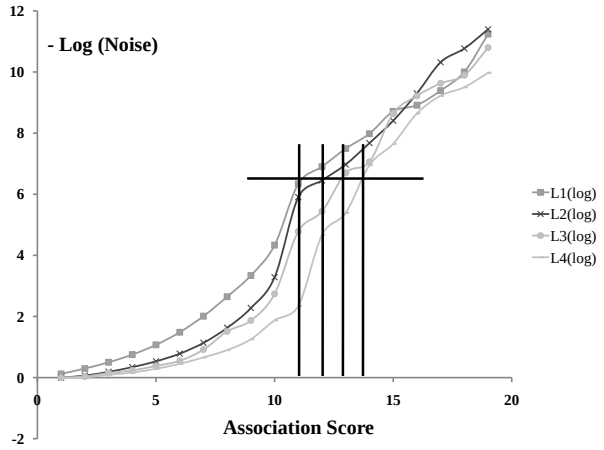

Figure 4: Noise as a function of the association score for the four complexity classes. We show the four curves in terms of $-log(Noise)$. A Noise level of 0.0015 (having -log equal to 6.5) indicated by the short horizontal line corresponds to different association scores in the four curves, roughly 11.0 for L^1, 12.0 for L^2, 13.0 for L^3 and 14.0 for L^4.

below, we illustrate the case of the $U^4 - g4$ table, where the horizontal axis corresponds to the number of bi-phrases that are kept, and the vertical axis to the BLEU score that is obtained by the translation system (Papineni et al., 2001). We see that by keeping only the 100000 bi-phrases with the highest association scores, we obtain performance which is almost indistinguishable from the performances obtained by keeping up to 350000 bi-phrases.

Second experiment: Pruning based on Noise with several complexity classes We now move to experiments where we actually use Noise as the pruning criterion, along with several complexity classes for computing it from raw association scores. In these experiments we take as our global bi-phrase table the table $U^4 - g0$, corresponding to bi-phrases obtained by combinations of up to 4 cepts, and containing no gap. We partition $U^4 - g0$ into 4 subtables $L^1 - g0$, $L^2 - g0$, $L^3 - g0$, $L^4 - g0$, corresponding to bi-phrases containing 1, 2, 3, and 4 cepts. We then compute four Noise curves relating the raw association score with the Noise level, one curve for each of $L^1 - g0$, $L^2 - g0$, $L^3 - g0$, $L^4 - g0$, as explained in section 4. Figure 4 represents the curves corresponding to $L^1 - g0$, $L^2 - g0$, $L^3 - g0$, $L^4 - g0$. Now that we have the four curves, we can experiment with different Noise levels, and see what are the consequences on the severity of pruning and on the translation performance. At the end of this process, we find that the optimal translation performance is obtained for a Noise of 0.0015, which corresponds approximately to the four association score thresholds $\beta 1 = 11.0$ for $L^1 - g0$, $\beta 2 = 12.0$ for $L^2 - g0$, $\beta 3 = 13.0$ for $L^3 - g0$, $\beta 4 = 14.0$ for $L^4 - g0$. Using this Noise, the pruned table size and translation performance results are as shown in Table 2. In Table 3, we contrast this with using either one of the association score thresholds for filtering the global table. From these tables, we draw two conclusions:

L^{1234} (309,908 bi-phrase)			
Threshold	bi-phrases	NIST	BLEU
11,12,13,14	108,159	6.8187	0.3749

Table 2: Pruned table size and translation performance when filtering the bi-phrases on a common Noise level of 0.0015, corresponding to using the different association score thresholds 11, 12, 13 and 14 on the four subtables respectively.

(1) When using the optimal common Noise level, we are able to prune about two thirds of the original bi-phrase table, and this would be roughly true also if we used either of the four association scores 11, 12, 13 or 14 as the sole criterion for pruning the table.

(2) However, none of the association score thresholds 11.0, 12.0, 13.0, 14.0, when taken as the sole criterion for pruning the table, obtains as good results (by some substantial margin relative to BLEU

$U^4 - g0$ (309,908 bi-phrase)			
Threshold	bi-phrases	NIST	BLEU
11	126,667	6.7485	0.3680
12	119,569	6.7460	0.3657
13	112,135	6.7263	0.3652
14	102,461	6.7492	0.3663

Table 3: Pruned table size and translation performance when filtering the bi-phrases on either one of the four association thresholds 11, 12, 13 or 14.

as well as NIST) as using the common Noise level of 0.0015, corresponding to applying the four association scores separately to the four complexity classes. Thus we see an experimental confirmation of our original intuition that computing the Noise without distinguishing between bi-phrases of different complexity can indeed be a suboptimal approach.

6 Conclusion

This paper has shown how the consideration of different complexity classes in a bi-phrase table has the effect of dissociating Noise from *p-value*, and that using Noise as the filtering criterion is superior to using *p-value* in term of translation (automatic) evaluation, while being comparable in terms of pruning strength. We have introduced a powerful and flexible way of computing Noise through simulation. While here we have explored only one way of stating the null hypothesis that neutralizes the translation relation in the bilingual corpus, in future work, we plan to explore some of the other variants that are made possible by our approach. As was also mentioned, there is an active subfield of Statistics with clear connections with the problem we investigated, Multiple Hypotheses Testing. Measures like the False Discovery Rate (Benjamini and Hochberg, 1995), and the methods for controlling it that have been proposed since, constitute additional promising directions for even more accurate phrase-table filtering.

Acknowledgments

This work was supported by the European Commission under the IST Project SMART (FP6-033917). Thanks to Eric Gaussier for his support at the beginning of this project, and to Sara Stymne and the anonymous reviewers for detailed and insightful comments.

References

Alan Agresti. 1992. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2007. Translation model pruning via usage statistics for statistical machine translation. In Candace L. Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, *HLT-NAACL (Short Papers)*, pages 21–24. The Association for Computational Linguistics.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL)*, Prague, June. Association for Computational Linguistics.

Adam Lopez. 2008. Statistical machine translation. *ACM Comput. Surv.*, 40(3):1–49.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 333–340, Barcelona, Spain, July. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. In *Proc. ACL-02*, pages 311–318, Morristown, NJ, USA.

R-Manual. 2009. Fisher's exact test for count data. Package stats version 2.7.0.

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with non-contiguous phrases. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 755–762, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Branko Soric. 1989. Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, 84(406):608–610, June.