Phrase Translation Model Enhanced with Association based Features

Boxing Chen, George Foster and **Roland Kuhn**
National Research Council
283 Taché Blvd. Gatineau (Quebec), Canada J8X 3X7
{Boxing.Chen, George.Foster, Roland.Kuhn}@nrc.ca

Abstract

In this paper, we propose to enhance the phrase translation model with association measures as new feature functions. These features are estimated on counts of phrase pair co-occurrence and their marginal counts. Four feature functions, namely, Dice coefficient, log-likelihood-ratio, hyper-geometric distribution and link probability are exploited and compared. Experimental results demonstrate that the performance of the phrase translation model can be improved by enhancing it with these association based feature functions. Moreover, we study the correlation between the features to predict the usefulness of a new association feature given the existing features.

1 Introduction

Phrase-based translation is one of the dominant current approaches to statistical machine translation (SMT). A phrase translation model, incorporated in a data structure known as a phrase table, is the most important component of a phrase-based SMT system, since the translations are generated by concatenating target phrases stored in the phrase table. The pairs of source and corresponding target phrases are extracted from the word-aligned bilingual training corpus (Och and Ney, 2003; Koehn et al., 2003). These phrase pairs together with useful feature functions, are called collectively a phrase translation model or phrase table.

A phrase translation model embedded in a state-of-the-art phrase-based SMT system normally exploits the feature functions involving conditional translation probabilities and lexical weights (Koehn et al., 2003). The phrase-based conditional translation probabilities are estimated from the relative frequencies of the source and target phrase in a given phrase pair. To avoid over-training, lexical weights are used to validate the quality of a phrase translation pair. They check how well the words in the phrase pair translate to each other. Such a phrase translation model has demonstrated its success in phrase-based SMT.

Zhao et al. (2004) improved the phrase table by using models employing *tf.idf* term weights to estimate the similarity between the source-language and target-language in a phrase pair; this measure will, for instance, assign higher similarity to a phrase pair if both of them contain a very rare word (in their respective languages). The similarity measure is used as an additional feature. The translation probability can be also discriminatively trained as in (Tillmann and Zhang, 2006; Liang et al., 2006). Deng et al. (2008) improved the algorithm of phrase extraction. To balance precision and recall, an information retrieval process implemented with a log-linear model is used to formulate the phrase pair extraction problem.

The quality of a phrase-pair can also be checked by other measures, such as measures of source and target phrase co-occurrence; we will call these association scores. Johnson et al. (2007) used phrase pair association based significance testing to discard most of the phrase pairs in a phrase table without loss of performance. However, if one discards phrase pairs from a phrase table, one must heuristically choose a threshold on the value of the association as a criterion for discarding.

Chen et al. (2005) used word-based association scores as rescoring features to re-rank the *n*-best lists of translations. The improvement reported in (Chen et al., 2005) confirmed that association based features are helpful for improving the translation performance. However, this work used the features for re-ranking rather than for decoding. Furthermore, the association features were based on words rather than on phrases, even though the latter is the basic translation unit.

Smadja et al. (1996) used association scores to produce translations for the collocations in one language from parallel corpora. This also proved that association scores are a good measure for phrase pairs in parallel corpora.

In this paper, we propose to enhance the phrase translation model with phrase pair co-occurrence based association scores as new feature functions. Association features are estimated from the counts of how often a given source-language phrase is in a training corpus sentence that is aligned with a sentence containing a given target-language phrase, and also on the overall counts of the source phrase and the target phrase in this corpus. Therefore, they can provide complementary information to the existing link-based estimates, such as phrase translation probabilities and lexical weights. Also, compared with discriminatively-learned scores as introduced in (Tillmann and Zhang, 2006; Liang et al., 2006), association features are easy to calculate.

We systematically explore the subset of phrase features that can be derived from standard association measures. We first exploit and compare four widely used feature functions, namely, Dice coefficient, log-likelihood-ratio, hyper-geometric distribution and link probability. We demonstrate that the performance of the phrase translation model can be improved by incorporating such feature functions. Moreover, we study the correlation between features to predict the usefulness of a new feature given the existing features, and thus introduce the technique of using correlation as a possible cheap way to search for useful new features.

This paper is organized as follows. Section 2 reviews the basic idea of a standard phrase-based SMT system. Section 3 proposes the idea of enhancing the translation model with new association-based features. Section 4 describes the experimental setting and results on NIST Chinese-to-English translation tasks. Section 5 discusses and analyzes the results obtained. And Section 6 concludes the paper.

2 SMT Baseline

Phrase-based statistical machine translation systems are usually modeled through a log-linear framework (Och and Ney, 2002) by introducing the hidden word alignment variable a (Brown et al., 1993).

$$\tilde{e}^{*} = \arg\max_{e,a}\left(\sum_{m=1}^{M} \lambda_m H_m(\tilde{e}, \tilde{f}, a)\right) \qquad (1)$$

where $\tilde{e}$ is a string of phrases in the target language, $\tilde{f}$ is the source language string, $H_m(\tilde{e}, \tilde{f}, a)$ are feature functions, and weights λ_m are typically optimized to maximize the scoring function (Och, 2003).

The SMT system we applied in our experiments is Portage, a state-of-the-art phrase-based translation system (Ueffing et al., 2007). The models and feature functions which are employed by the decoder are:

- Phrase translation model(s) estimated on the word alignments generated using IBM model 2 and HMMs (Vogel et al., 1996), with standard phrase conditional translation probabilities and lexical weights being employed;
- Distortion model, which assigns a penalty based on the number of source words which are skipped when generating a new target phrase;
- Language model(s) trained using SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing method (Chen et al, 1996);
- Word/phrase penalties.

3 Phrase Translation Model with Association-based Features

3.1 Traditional phrase-table features

A typical phrase translation model exploits features estimating phrase conditional translation probabilities and lexical weights. The phrase conditional translation probabilities are estimated from relative frequencies. Normally, two-directions, i.e., direct and inverse phrase translation probabilities are exploited. The direct phrase-based translation probability is computed as follows:

$$\mathrm{P}(\tilde{f}|\tilde{e}) = \frac{link(\tilde{f},\tilde{e})}{\sum_{\tilde{f}'} link(\tilde{f}',\tilde{e})} \qquad (2)$$

where $link(\tilde{f},\tilde{e})$ is the frequency of source phrase $\tilde{f}$ and target phrase $\tilde{e}$ linked to each other.

Most of the longer phrases are seen only once in the training corpus. Therefore, relative frequencies overestimate the phrase translation probabilities. To overcome the overestimation problem, lexical weights are introduced.

There are various approaches for estimating the lexical weights. The first approach is intro-

duced in (Koehn et al., 2003), Given a phrase pair $(\tilde{f}, \tilde{e})$ and a word alignment a between the source word position $i = 1, \ldots, n$ and the target word position $j = 0,1, \ldots, m$, the lexical weight is computed by

$$P_w(\tilde{f}|\tilde{e}) = \prod_{i=1}^{n} \frac{1}{|\{j|(i.j)\in a\}|} \sum_{\forall(i,j)\in a} p(f_i|e_j) \quad (3)$$

where $p(f_i|e_j)$ is the lexical translation probability, which is computed from the relative frequencies in the same word-aligned training corpus.

The second approach calculates lexical weights based on a statistical lexicon for the constituent words in the phrase (Vogel et al., 2003).

$$P_w(\tilde{f}|\tilde{e}) = \prod_{i=1}^{n} \sum_{j=1}^{m} p(f_i|e_j) \quad (4)$$

where $p(f_i|e_j)$ is word probability estimated using IBM alignment model 1, without considering the position alignment probabilities.

The third approach assumes that all source words are conditionally independent, so that:

$$P_w(\tilde{f}|\tilde{e}) = \prod_{i=1}^{n} p(f_i|\tilde{e}) \quad (5)$$

To compute $p(f_i|\tilde{e})$, Zens and Ney (2004) describe a "noisy-or" combination:

$$p(f_i|\tilde{e}) = 1 - p(\neg f_i|\tilde{e}) \approx 1 - \prod_{j=1}^{m}(1 - p(f_i|e_j)) \quad (6)$$

where $p(\neg f_i|\tilde{e})$ is the probability that f_i is not in the translation of $\tilde{e}$, and $p(f_i|e_j)$ is a lexical probability. Zens and Ney (2004) obtain the lexical probability from smoothed relative-frequency estimates from the same word aligned training corpus. Foster et al. (2006) uses IBM1 probabilities to perform further smoothing.

3.2 Additional phrase-table association features

On top of these traditional feature functions which are based on the phrase-pair link count, we propose to exploit additional feature functions for the phrase table based on phrase-pair co-occurrence information. All of these consider a co-occurrence of a source phrase and a target phrase to have occurred if they are found in sentences that have been aligned with each other (no alignment between the words in each is necessary). We study four commonly used association-based features, namely, Dice coefficient, log-likelihood-ratio, hyper-geometric distribution, and link probability in this paper.

	$\tilde{f}$	$\neg\tilde{f}$	
$\tilde{e}$	$C(\tilde{f},\tilde{e})$	$C(\neg\tilde{f},\tilde{e})$	$C(\tilde{e})$
$\neg\tilde{e}$	$C(\tilde{f},\neg\tilde{e})$	$C(\neg\tilde{f},\neg\tilde{e})$	$C(\neg\tilde{e})$
	$C(\tilde{f})$	$C(\neg\tilde{f})$	N

Table 1: Contingency table for phrase $\tilde{f}$ and $\tilde{e}$.

Let $C(\tilde{f},\tilde{e})$ denotes the co-occurrence count[1] of source phrase $\tilde{f}$ and target phrase $\tilde{e}$ in the parallel training corpus. It is worth noting that phrase $\tilde{f}$ and phrase $\tilde{e}$ are not necessarily linked to each other. $C(\tilde{f})$ and $C(\tilde{e})$ denote the count of $\tilde{f}$ in the source sentences and the count of $\tilde{e}$ in the target sentences, respectively. $\neg\tilde{f}$ and $\neg\tilde{e}$ mean that corresponding phrases do not occur in the respective sentences. *N* represents the total number sentence pairs in the parallel training corpus. For each phrase, only one occurrence per sentence is taken into account. The statistics can be organized in a contingency table, e.g. in Table 1. When collecting the statistics of the data, we only need to count $C(\tilde{f},\tilde{e})$, $C(\tilde{f})$, $C(\tilde{e})$ and *N*; the other counts could be easily calculated accordingly. Then, we may compute the following association features:

1) **Dice coefficient** (Dice, 1945) as in Equation (7). It compares the co-occurrence count of phrase pair $\tilde{f}$ and $\tilde{e}$ with the sum of the independent occurrence counts of $\tilde{f}$ and $\tilde{e}$.

$$Dice(\tilde{f},\tilde{e}) = \frac{2\times C(\tilde{f},\tilde{e})}{C(\tilde{f})+C(\tilde{e})} \quad (7)$$

2) **Log-likelihood-ratio** (Dunning, 1993) as in Equation (8) which is presented by Moore (2004).

$$LLR(\tilde{f},\tilde{e}) = \sum_{\tilde{f}?\in\{\tilde{f},\neg\tilde{f}\}}\sum_{\tilde{e}?\in\{\tilde{e},\neg\tilde{e}\}} C(\tilde{f}?,\tilde{e}?) log \frac{N\times C(\tilde{f}?,\tilde{e}?)}{C(\tilde{f}?)\times C(\tilde{e}?)} \quad (8)$$

where $\tilde{f}?$ and $\tilde{e}?$ are variables ranging over the values $\{\tilde{f},\neg\tilde{f}\}$ and $\{\tilde{e},\neg\tilde{e}\}$ respectively, $C(\tilde{f}?,\tilde{e}?)$ is the joint count for the values of $\tilde{f}?$ and $\tilde{e}?$, $C(\tilde{f}?)$ and $C(\tilde{e}?)$ are the frequencies of values of $\tilde{f}?$ and $\tilde{e}?$.

3) **Hyper-geometric distribution** is the probability of the phrase-pair globally co-occurring $C(\tilde{f},\tilde{e})$ times among *N* sentence-pairs by chance, given the marginal frequencies $C(\tilde{f})$ and $C(\tilde{e})$. We computed this prob-

[1] There are at least three ways to count the number of co-occurrence of $\tilde{f}$ and $\tilde{e}$, if one or both of them have more than one occurrence in a given sentence pair (Melamed, 1998). We choose to count the co-occurrence as 1 if they both occur; otherwise 0.

ability assuming a binomial distribution. Without simplification, this probability can be expressed by:

$$P_{hg}\left(C(\tilde{f},\tilde{e})\right) = \frac{\binom{C(\tilde{f})}{C(\tilde{f},\tilde{e})}\binom{C(\neg\tilde{f})}{C(\neg\tilde{f},\tilde{e})}}{\binom{N}{C(\tilde{e})}} \quad (9)$$

However, because this probability is very small, we follow (Johnson et al., 2007) in computing the negative of the natural logs of the p-value associated with the hypergeometric distribution as the feature functions:

$$H_{hg}(\tilde{f},\tilde{e}) = -\log\left(\sum_{k=C(\tilde{f},\tilde{e})}^{\infty} P_{hg}(k)\right) \quad (10)$$

Therefore, the higher the value of $H_{hg}(\tilde{f},\tilde{e})$, the stronger the association between phrase $\tilde{f}$ and phrase $\tilde{e}$.

4) **Link probability** (Moore, 2005) is the conditional probability of the phrase pair being linked given that they co-occur in a given sentence pair. It is estimated as:

$$LP\left(\tilde{f},\tilde{e}\right) = \frac{link(\tilde{f},\tilde{e})}{C(\tilde{f},\tilde{e})} \quad (11)$$

where $link(\tilde{f},\tilde{e})$ is the number of times that $\tilde{f}$ and $\tilde{e}$ are linked in sentence pairs.

3.3 Phrase-table smoothing

As (Foster et al., 2006) shows, the phrase table can be improved by applying smoothing techniques. A motivation for this is our observation that the phrase pairs which co-occur only once: $C(\tilde{f},\tilde{e}) = 1$ are amazingly frequent in the phrase table even when the training corpus is very large. To compensate for this over-confidence in rare events, we apply absolute discounting.

$$C_d(\tilde{f},\tilde{e}) = C(\tilde{f},\tilde{e}) - d_1 \quad (12)$$
$$C_d(\tilde{f}) = C(\tilde{f}) - d_2 \quad (13)$$
$$C_d(\tilde{e}) = C(\tilde{e}) - d_2 \quad (14)$$

The optimal values of d_1 and d_2 are determined by heuristic tuning on the dev set.

4 Experiments

4.1 Experimental setting

Since the values of association scores are strongly related to the corpus size, i.e., the total number of events, we carried out the experiments in two data conditions. The first one is the *large data* condition which is distributed for NIST[2] 2008 evaluation Chinese-to-English track. In particular, all the allowed bilingual corpora except the UN corpus have been used for estimating the phrase table. The second one is the *small data* condition where only the FBIS[3] corpus is used to train the phrase-table. However, we used the same two language models for both of these conditions. The first one is a 5-gram LM which is estimated on the target side of the texts used in the *large data* condition. The second LM is a 5-gram LM trained on the so-called English Gigaword corpus, one of the allowed monolingual resources for the NIST task.

We carried out experiments for translating Chinese to English. We use the same development and test sets for the two data conditions. We first created a development set which used mainly data from the NIST 2005 test set, and also some balanced-genre web-text from the NIST training material. Evaluation was performed on the NIST 2004, 2006 and 2008 test sets. Table 2 gives figures about training, development and test corpora; |S| is the number of the sentences, and |W| is the size of running words.

In our baseline system, we applied smoothing technique "RF+ZN-IBM1" proposed by Foster et al. (2006) to estimate the phrase translation probabilities and lexical weights. This smoothing technique achieved the best performance in our experiments which was consistent with the results reported in (Foster et al., 2006).

			Chi	Eng
Parallel Train	Large Data	|S|	3,775K	
		|W|	111.2M	112.6M
	Small Data	|S|	245K	
		|W|	9.0M	10.5M
Dev		|S|	1,500	1,500×4
		|W|	38K	160K
Test	NIST04	|S|	1,788	1,788×4
	NIST06	|S|	1,664	1,664×4
	NIST08	|S|	1,357	1,357×4
Gigaword		|S|	-	11.7M

Table 2: Statistics of training, dev, and test sets.

4.2 Results

Our evaluation metric is IBM BLEU (Papineni et al., 2002), which performs case-insensitive matching of *n*-grams up to *n* = 4.

[2] http://www.nist.gov/speech/tests/mt
[3] LDC2003E14

First, the results without performing count smoothing on association features are reported in Table 3. The row LD shows the results of large data track, and SD for small data track. Δ is the average improvement of BLEU-score over the baseline for three test sets. Note that three association features, namely, Dice coefficient ("dc"), log-likelihood-ratio ("llr"), and hyper-geometric distribution ("hg") improved the performance on all three test sets for both data tracks. However, link probability ("lp") failed to give noticeable and consistent improvement. Log-likelihood-ratio showed the biggest improvement (0.44 BLEU-point on average for three test sets on the large data condition and 0.52 BLEU-point on small data condition), the runner-up was Dice coefficient (about 0.32 BLEU-point on large data and 0.37 on small data).

system		2004	2006	2008	Δ
LD	baseline	34.83	30.95	24.91	-
	+dc	35.20	31.24	25.21	0.32
	+llr	**35.28**	**31.38**	**25.36**	**0.44**
	+hg	35.13	31.06	25.16	0.22
	+lp	34.93	30.98	24.85	0.02
SD	baseline	30.88	27.07	21.37	-
	+dc	31.21	27.44	21.78	0.37
	+llr	**31.39**	**27.59**	**21.89**	**0.52**
	+hg	31.19	27.31	21.73	0.30
	+lp	30.79	27.05	21.22	-0.08

Table 3: Results (BLEU% score) of systems without counts smoothing. "dc" is Dice coefficient; "llr" is log-likelihood-ratio; "hg" is hyper-geometric distribution; and "lp" is link probability. LD is large data track, and SD is small data track. Δ is the average improvement of BLEU-score over the baseline for three test sets.

System		2004	2006	2008	Δ
LD	baseline	34.83	30.95	24.91	-
	+dc	35.30	31.29	25.26	0.38
	+llr	35.29	**31.39**	**25.41**	**0.47**
	+hg	**35.35**	31.38	25.33	0.46
	+lp	34.89	31.01	24.95	0.05
SD	baseline	30.88	27.07	21.37	-
	+dc	31.42	27.64	21.97	0.57
	+llr	31.47	**27.69**	**22.08**	**0.64**
	+hg	**31.51**	27.55	21.96	0.56
	+lp	30.90	27.04	21.27	-0.04

Table 4: Results (BLEU% score) of systems with counts smoothing.

For count smoothing, first we need to set the values of d_1 and d_2 for each association feature. By observing translation performance obtained from the additional experiments on the development set, we heuristically determined the optimal values of d_1 and d_2 for Dice coefficient and hyper-geometric distribution are 0.9 and 0.8 respectively. The optimal values of d_1 and d_2 for log-likelihood-ratio are 0.1 for both, on both data conditions. The results with count smoothing are reported in Table 4. The first three features still obtained improvement compared to the baseline. If one compares the results to those without count smoothing, on large data condition, the improvements are very small for log-likelihood-ratio and Dice coefficient; the improvements are more noticeable for hyper-geometric distribution, such as the 0.32 improvement in BLEU-score (from 31.06 to 31.38) that has been observed on NIST06 test set. On the small data track, the improvements produced by the counts smoothing are more similar; for Dice coefficient and hyper-geometric distribution, the further improvements are about 0.2 BLEU-score, and 0.1 BLEU-score for log-likelihood-ratio. In total, the improvements over the baseline for the first three features are quite similar. The range is from 0.38 (dc) to 0.47 (llr) BLEU-score for large data, and from 0.56 (hg) to 0.64 (llr) for small data.

We then study the combination of association features. Because the link probability has not shown good improvement, we only experimented with combinations of the first three features. Table 5 illuminates the results for combinations of three features. Combinations of two features further improved the performance. "+dc+hg", and "+llr+hg",and "+dc+llr" are all improvements over using a single association feature. However, a combination of three features failed to obtain further improvement.

System		2004	2006	2008	Δ
LD	baseline	34.83	30.95	24.91	-
	+dc+hg	**35.41**	**31.59**	25.59	**0.63**
	+dc+llr	35.38	31.49	25.43	0.53
	+llr+hg	35.39	31.57	**25.60**	0.62
	+dc+llr+hg	35.38	31.55	25.53	0.59
SD	baseline	30.98	27.17	21.47	-
	+dc+hg	31.62	27.75	22.19	0.64
	+dc+llr	**31.71**	27.79	22.10	0.66
	+llr+hg	31.66	**27.89**	**22.23**	**0.72**
	+dc+llr+hg	31.51	27.80	22.23	0.64

Table 5: Results (BLEU% score) of systems for the combinations of association features with counts smoothing.

1	reference	torrential rains hit western india , 43 people dead
	baseline	western india rains flooded 43 people were killed
	+dc+hg	western india 's heavy rains flooded 43 people were killed
2	reference	the emergency government 's authority expires today .
	baseline	an emergency government 's mandate expires today .
	+dc+hg	an emergency government 's authority expires today .
3	reference	" who are you looking for ? "
	baseline	" who you find ?
	+dc+hg	" who are you looking for ?
4	reference	this system will also be gradually rolled out nationwide .
	baseline	this system will gradually open the country .
	+dc+hg	this system will also gradually spread throughout the country .
5	reference	so , any mistake will lead to serious consequences .
	baseline	well , any error that will lead to serious consequences .
	+dc+hg	well , any mistake will lead to serious consequences .

Figure 1: Examples of the translations of baseline and "+dc+hg" system on large data condition.

	p_2	p_3	p_4	dc	llr	hg	lp
p_1	0.2510	0.4914	-0.4148	-0.1063	-0.1659	-0.2896	**0.7016**
p_2	-	-0.0078	0.1327	-0.3493	0.1942	-0.1367	0.0586
p_3	-0.0078	-	-0.3156	0.2748	-0.2390	-0.3237	**0.5725**
p_4	0.1327	-0.3156	-	0.0972	0.3729	0.3167	-0.3462
dc	-0.3493	0.2748	0.0972	-	0.4615	0.3733	-
llr	0.1942	-0.2390	0.3729	0.4615	-	**0.7969**	-

Table 7: Correlation coefficients between features for large data condition, where p_i (i=1,2,3,4) are traditional phrase-table features: direct phrase-based translation probability, direct lexical weights, inverse phrase-based translation probability, inverse lexical weights. "lp" is link probability.

	p_2	p_3	p_4	dc	llr	hg	lp
p_1	0.3564	0.3732	-0.1973	-0.1108	-0.0378	-0.2591	**0.6974**
p_2	-	-0.0074	0.0344	-0.2979	0.2257	-0.0831	0.1171
p_3	-0.0074	-	-0.1388	0.1483	-0.1751	-0.2755	**0.5227**
p_4	0.0344	-0.1388	-	0.0152	0.2185	0.2561	-0.1484
dc	-0.2979	0.1483	0.0152	-	**0.5970**	0.4031	-
llr	0.2257	-0.1751	0.2185	**0.5970**	-	**0.7885**	-

Table 8: Correlation coefficients between features for small data condition.

If we take "+dc+hg" as the optimal system, the improvements over the baseline for large data condition are 0.58 BLEU-score for NIST04, 0.64 BLEU-score for NIST06, and 0.68 BLEU-score for NIST08; for small data condition, if we take "+llr+hg" as the optimal system, the improvements over the baseline are 0.68 BLEU-score for NIST04, 0.72 BLEU-score for NIST06, and 0.76 BLEU-score for NIST08.

5 Discussion

Association features can improve the performance because they provide additional information about how well phrase pairs translate to each other, beyond that yielded by phrase-based conditional probabilities and lexical weights. Figure 1 shows how this additional information may be applied, using some typical examples which are selected from the test sets.

Consider the first example. The only difference between the baseline output and the improved output is *rains* and *heavy rains*, they are translations of the Chinese phrase 豪 雨 (pronounces "*háo yǔ*" and meaning "*torrential rains*"). According to the reference, *heavy rains* should be the better translation. However, the phrase translation conditional probabilities and lexical weights of these two phrase pairs are am-

biguous. Table 6 gives values of the two association features; the target phrase *heavy rains* has higher value than the phrase *rains* on both Dice coefficient and hyper-geometric distribution features. This could help the decoder choose *heavy rains* as the final translation for the source phrase.

豪 雨	Dice	H_{hg}
rains	0.1666	11.67
heavy rains	0.1995	22.93

Table 6: Dice coefficient and hyper-geometric distribution scores of target phrases *rains* and *heavy rains* given source phrase *豪 雨*.

Nonetheless, more association features cannot guarantee further improvement. The reason is that some association features are highly correlated with each other, since all of them are based on the co-occurrence and marginal counts. We study this issue via *Pearson product-moment correlation coefficient* (Rodgers and Nicewander, 1988) which measures the strength of a linear relationship between two features. We use the Equation (15) to compute the correlation coefficient.

Suppose we have a series of *n* values of two features H_1 and H_2 written as x_i and y_i (where $i = 1, 2, \ldots, n$), then the correlation coefficient can be used to estimate the correlation of H_1 and H_2. The correlation coefficient is written as:

$$R(H_1, H_2) = \frac{\sum_{i=0}^{i=n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} \qquad (15)$$

where $\bar{x}$ and $\bar{y}$ are the sample means of H_1 and H_2, s_x and s_y are the sample standard deviations of H_1 and H_2. The correlation coefficient ranges from -1 to 1. The closer the coefficient is to either -1 or 1, the stronger the correlation between the features.

We randomly select 1,000 phrase pairs from the phrase table, and then compute the pair-wise correlation coefficients for the following features:

a. Four traditional phrase-table futures;
b. Four association features introduced in section 3.2.

Table 7 and 8 list the correlation coefficients between all the features, where $p_i (i = 1,2,3,4)$ are traditional phrase-table features: direct phrase-based translation probability, direct lexical weights, inverse phrase-based translation probability, and inverse lexical weights. Please note that a logarithm operation is applied to all the variables before they are input to compute the correlation coefficients.

Not surprisingly, link probability is highly correlated with the phrase-based direct and inverse probability, since all these three scores are conditioned on the link number $link(\tilde{f}, \tilde{e})$. Moreover, we can also notice that link probability cannot guarantee an improvement for translation performance. The reason for this is that the information contained in the link probability has already been provided by either direct phrase-based translation probability or inverse phrase-based translation probability.

On the other hand, we can see that no original phrase-table features are highly correlated with Dice coefficient, log-likelihood-ratio, or hyper-geometric distribution (the correlation coefficient of greatest magnitude is 0.3729 which is the correlation between log-likelihood-ratio and inverse lexical weights p_4). Thus, these three association features can provide additional information about the phrase-pair to the decoder, improving the translation performance. However, these three features are correlated with each other; in particular, the log-likelihood-ratio is highly correlated with the hyper-geometric distribution. From the experiments involving feature link probability, we see that it is difficult to obtain further improvement if the new feature is highly correlated with existing features. This could explain why exploiting three features together does not yield an improvement over using only two of them.

6 Conclusions

In this paper, we proposed a way of using phrase pair association measures as new feature functions to enhance the translation model. These features are estimated from counts of phrase pair co-occurrences and the counts of the phrases themselves. We exploited and compared four well known and widely used feature functions: the Dice coefficient, log-likelihood-ratio, hyper-geometric distribution, and link probability. Experiments were carried out on NIST tasks on large and small data conditions. The results on the two tasks both demonstrated that the translation model will perform better if the association features are integrated into a standard phrase table. Moreover, we studied the correlation between the features to predict the usefulness of a new feature given the existing features. We learned that it is difficult to obtain further improvement when applying a new feature which is highly correlated with existing features, thus for the technique of using correlation as a possible cheap way to search for useful new features.

Acknowledgments
This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

P. F. Brown, V. J. Della Pietra, S. A. Della Pietra and R. L. Mercer. 1993. The Mathematics of Statistical Ma-chine Translation: Parameter Estimation. *Computational Linguistics*, 19(2) 263-312.

B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo and M. Federico. 2005. The ITC-irst SMT System for IWSLT-2005. In *Proceeding of IWSLT-2005*, pp.98-104, Pittsburgh, USA, October.

S. F. Chen and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL*, pages 310–318.

Y. Deng, J. Xu, and Y, Gao. 2008. Phrase Table Training for Precision and Recall: What Makes a Good Phrase and a Good Phrase Pair? In *Proceedings of ACL*, pp 81-88.

L. R. Dice. 1945. Measures of the amount of ecologic association between species. *Journal of Ecology*, 26:297–302.

T. Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

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

H. Johnson, J. Martin, G. Foster, and R. Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL*, pages 967–975.

P. Koehn, F. J. Och, D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL 2003*.

P. Liang, A. Bouchard-Cote, D. Klein, and B. Taskar. 2006. An End-to-End Discriminative Approach to Machine Translation. In *Proceedings of COLING-ACL 2006*.

I. D. Melamed. 1998. Models of Co-occurrence. *University of Pennsylvania, IRCS Technical Report #98-05.*

R. 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*, pp. 333–340, Barcelona, Spain.

R. C. Moore. 2005. Association-based bilingual word alignment. In *Proceedings of Workshop of SMT of ACL* 2005, pp1-8.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL-2003*. Sapporo, Japan.

F. J. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceeding of ACL-2002*.

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

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

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

F. Smadja, K. McKeown, and V. Hatzivassiloglou, 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics.* 22(1), 1-38.

A. Stolcke. 2002. SRILM - an extensible language modelling toolkit. In *Proceeding of ICSLP-2002*. 901-904.

C. Tillmann and T. Zhang. 2006. A discriminative global training algorithm for statistical MT. In *Proceedings of ACL*, pages 721–728.

N. Ueffing, M. Simard, S. Larkin, and J. H. Johnson. 2007. NRC's Portage system for WMT 2007. In *Proceedings ACL Workshop on SMT*.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM based word alignment in statistical translation. In *Proceedings of the COLING*.

S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venogupal, B. Zhao, A. Waibel. 2003. The CMU Statistical Translation System. In *Proceedings of MT-Summit IX*. New Orleans, LA.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proceedings of Human Language Technology Conference / North American Chapter of the ACL*, Boston, May.

B. Zhao, S. Vogel, M. Eck, and A. Waibel. 2004. Phrase pair rescoring with term weighting for statistical machine translation. In *Proceedings of EMNLP*, pages 206–213.