Feature-Rich Discriminative Phrase Rescoring for SMT

Fei Huang and **Bing Xiang**

IBM T. J. Watson Research Center
{huangfe, bxiang}@us.ibm.com

Abstract

This paper proposes a new approach to phrase rescoring for statistical machine translation (SMT). A set of novel features capturing the translingual equivalence between a source and a target phrase pair are introduced. These features are combined with linear regression model and neural network to predict the quality score of the phrase translation pair. These phrase scores are used to discriminatively rescore the baseline MT system's phrase library: boost good phrase translations while prune bad ones. This approach not only significantly improves machine translation quality, but also reduces the model size by a considerable margin.

1 Introduction

Statistical Machine Translation (SMT) systems, including phrase-based (Och and Ney 2002; Koehn et. al. 2003), syntax-based (Yamada and Knight 2001; Galley et. al. 2004) or hybrid systems (Chiang 2005; Zollmann and Venugopal 2006), are typically built with bilingual phrase pairs, which are extracted from parallel sentences with word alignment. Due to the noises in the bilingual sentence pairs and errors from automatic word alignment, the extracted phrase pairs may contain errors, such as

- dropping content words
  (the $num countries ,||个:<null>),
- length mismatch
  (along the lines of the || 的:of)
- content irrelevance
  (the next $num years, ||
  水平:level 方面:aspect 所:<null>)

These incorrect phrase pairs compete with correct phrase pairs during the decoding process, and are often selected when their counts are high (if they contain systematic alignment errors) or certain model costs are low (for example, when some source content words are translated into target function words in an incorrect phrase pair, the language model cost of the incorrect pair may be small, making it more likely that the pair will be selected for the final translation). As a result, the translation quality is degraded when these incorrect phrase pairs are selected.

Various approaches have been proposed over the past decade for the purpose of improving the phrase pair quality for SMT. For example, a term weight based model was presented in (Zhao, et al., 2004) to rescore phrase translation pairs. It models the translation probability with similarities between the query (source phrase) and document (target phrase). Significant improvement was obtained in the translation performance. In (Johnson, et al., 2007; Yang and Zheng, 2009), a statistical significance test was used to heavily prune the phrase table and thus achieved higher precision and better MT performance.

In (Deng, et al., 2008), a generic phrase training algorithm was proposed with the focus on phrase extraction. Multiple feature functions are utilized based on information metrics or word alignment. The feature parameters are optimized to directly maximize the end-to-end system performance. Significant improvement was reported for a small MT task. But when the phrase table is large, such as in a large-scale SMT system, the computational cost of tuning with this approach will be high due to many iterations of phrase extraction and re-decoding.

In this paper we attempt to improve the quality of the phrase table using discriminative phrase rescoring method. We develop extensive set of features capturing the equivalence of bilingual

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 492–500,
Beijing, August 2010

phrase pairs. We combine these features using linear and nonlinear models in order to predict the quality of phrase pairs. Finally we boost the score of good phrases while pruning bad phrases. This approach not only significantly improves the translation quality, but also reduces the phrase table size by 16%.

The paper is organized as follows: in section 2 we discuss two regression models for phrase pair quality prediction: linear regression and neural network. In section 3 we introduce the rich set of features. We describe how to obtain the training data for supervised learning of the two models in section 4. Section 5 presents some approaches to discriminative phrase rescoring using these scores, followed by experiments on model regression and machine translation in section 6.

2 Problem Formulation

Our goal is to predict the translation quality of a given bilingual phrase pair based on a set of features capturing their similarities. These features are combined with linear regression model and neural network. The training data for both models are derived from phrase pairs extracted from small amount of parallel sentences with hand alignment and machine alignment. Details are given in section 4.

2.1 Linear regression model

In the linear regression model, the predicted phrase pair quality score is defined as

$$Sco(e,f) = \sum_i \lambda_i f_i(e,f) \qquad (1)$$

where $f_i(e,f)$ is the feature for the phrase pair (*e*,*f*), as to be defined in section 3. These feature values can be binary (0/1), integers or real values. λ s are the feature weights to be learned from training data. The phrase pair quality score in the training data is defined as the sum of the target phrase's BLEU score (Papineni et. al. 2002) and the source phrase's BLEU score, where the reference translation is obtained from phrase pairs extracted from human alignment. Details about the training data are given in section 4. The linear regression model is trained using a statistical package R[1]. After training, the learned feature weights are applied on a held-out set of phrase pairs with known quality scores to evaluate the model's regression accuracy.

2.2 Neural Network model

A feed-forward back-propagation network (Bryson and Ho, 1969) is created with one hidden layer and 20 nodes. During training, the phrase pair features are fed into the network with their quality scores as expected outputs. After certain iterations of training, the neural net's weights are stable and its mean square error on the training set has been significantly reduced. Then the learned network weights are fixed, and are applied to the test phrase pairs for regression accuracy evaluation. We use MatLab™'s neural net toolkit for training and test.

We will compare both models' prediction accuracy in section 6. We would like to know whether the non-linear regression model outperforms linear regression model in terms of score prediction error, and if fewer regression errors correspond to better translation quality.

3 Feature Description

In this section we will describe the features we use to model the equivalence of a bilingual phrase pair (*e*,*f*). These features are defined on the phrase pair, its compositional units (words and characters), attributes (POS tags, numbers), co-occurrence frequency, length ratio, coverage ratio and alignment pattern.

- Phrase : $P_p(f \mid e)$, $P_p(e \mid f)$

$$P_p(e \mid f) = \frac{C(e,f)}{C(f)} \qquad (2)$$

where $C(e,f)$ is the co-occurrence frequency of the phrase pair (*e*,*f*), and C(*f*) is the occurrence frequency of the source phrase *f*. $P_p(f \mid e)$ is defined similarly.

- Word : $P_w(f \mid e)$, $P_w(e \mid f)$

$$P_w(e \mid f) = \prod_i \max_j t(e_i \mid f_j) \qquad (3)$$

where $t(e_i \mid f_j)$ is the lexical translation probability. This is similar to the word-level phrase

[1] http://www.r-project.org/

translation probability, as typically calculated in SMT systems (Brown et. al. 1993). Here we use max instead of sum. $P_w(f \mid e)$ is calculated similarly.

- Character: $P_c(f \mid e), P_c(e \mid f)$

When the source or target words are composed of smaller units, such as characters for Chinese words, or prefix/stem/suffix for Arabic words, we can calculate their translation probability on the sub-unit level. This is helpful for languages where the meaning of a word is closely related to its compositional units, such as Chinese and Arabic.

$$P_c(e \mid f) = \prod_i \max_n t(e_i \mid c_n) \qquad (4)$$

where c_n is the n-th character in the source phrase f (n=1,…,N).

- POS tag: $P_t(f \mid e), P_t(e \mid f)$

In addition to the probabilities estimated at the character, word and phrase levels based on the surface forms, we also compute the POS-based phrase translation probabilities. For each source and target word in a phrase pair, we automatically label their POS tags. Then POS-based probabilities are computed in a way similar to the calculation of the word-level phrase translation probability (formula 3). It is believed that such syntactic information can help to distinguish good phrase pairs from bad ones (for example, when a verb is aligned to a noun, its POS translation probability should be low).

- Length ratio

This feature computes the ratio of the number of content words in the source and target phrases. It is designed to penalize phrases where content words in the source phrase are dropped in the target phrase (or vice versa). The ratio is defined to be 10 if the target phrase has zero content word while the source phrase has non-zero content words. If neither phrase contains a content word, the ratio is defined to be 1.

- Log frequency

This feature takes the logarithm of the co-occurrence frequency of the phrase pair. High frequency phrase pairs are more likely to be correct translations if they are not due to systematic alignment errors.

- Coverage ratio

We propose this novel feature based on the observation that if a phrase pair is a correct translation, it often includes correct sub-phrase pair translations (*decomposition*). Similarly a correct phrase pair will also appear in correct longer phrase pair translations (*composition*) unless it is a very long phrase pair itself. Formally we define the coverage ratio of a phrase pair (*e*,*f*) as:

$$Cov(e, f) = Cov_d(e, f) + Cov_c(e, f) . \quad (5)$$

Here $Cov_d(e, f)$ is the decomposition coverage:

$$Cov_d(e, f) = \sum_{f_i \subseteq f} \frac{\sum_{(e_i, f_i) \in P_L} \Delta(e_i, e)}{\sum_{(*, f_i) \in P_L} 1} \qquad (6)$$

where f_i is a sub-phrase of f, and (e_i, f_i) is a phrase pair in the MT system's bilingual phrase library P_L. $\Delta(e_1, e_2)$ is defined to be 1 if $e_1 \subseteq e_2$, otherwise it is 0. For each source sub-phrase f_i, this formula calculates the ratio that its target translation e_i is also a sub-phrase of the target phrase *e*, then the ratio is summed over all the source sub-phrases.

Similarly the composition coverage is defined as

$$Cov_c(e, f) = \sum_{f \subseteq f^j} \frac{\sum_{(e^j, f^j) \in P_L} \Delta(e, e^j)}{\sum_{(*, f^j) \in P_L} 1} \qquad (7)$$

where f^j is any source phrase containing f and e^j is one of f^j's translations in P_L. We call f^j a super-phrase of *f*. For each source super-phrase f^j, this formula calculates the ratio that its target translation e^j is also a super-phrase of the target phrase *e*, then the ratio is summed over all the source super-phrases.

Short phrase pairs (such as a phrase pair with one source word translating into one target word) have less sub-phrases but more super-phrases (for long phrase pairs, it is the other way around).

Combining the two coverage factors produces balanced coverage ratio, not penalizing too short or too long phrases.

- Number match

During preprocessing of the training data, numbers are mapped into a special token ($num) for better generalization. Typically one number corresponds to one special token. During translation numbers should not be arbitrarily dropped or inserted. Therefore we can check whether the source and target phrases have the right number of $num to be matched. If they are the same the number match feature has value 1, otherwise it is 0.

- Alignment pattern

This feature calculates the number of *unaligned* content words in a given phrase pair, where word alignment is obtained simply based on the maximum lexical translation probability of the source (target) word given all the target (source) words in the phrase pair.

Among the above 13 features, the number match feature is a binary feature, the alignment pattern feature is an integer-value feature, and the rest are real-value features. Also note that most features are positively correlated with the phrase translation quality (the greater the feature value, the more likely it is a correct phrase translation) except the alignment pattern feature, where more unaligned content words corresponds to bad phrase translations.

4 Training Data

The training data for both the linear regression and neural network models are bilingual phrase pairs with the above 13 feature values as well as their expected phrase quality scores. The feature values can be computed according to the description in section 3. The expected translation quality score for the phrase pair (*e*,*f*) is defined as

$$B(e,f) = Bleu(e,e^* \mid f) + Bleu(f,f^* \mid e) \quad (8)$$

where e^* is the human translation of the source phrase *f*, and f^* is the human translation of the target phrase *e*. These human translations are obtained from hand alignment of some parallel sentences.

1. Given hand alignment of some bilingual sentence pairs, extract gold phrase translation pairs.
2. Apply automatic word alignment on the same bilingual sentences, and extract phrase pairs. Note that due to the word alignment errors, the extracted phrase pairs are noisy.
3. For each phrase pair (*e*, *f*) in the noisy phrase table, find whether the source phrase *f* also appears in the gold phrase table as (*e**, *f*). If so, use the corresponding target phrase(s) *e** as reference translation(s) to evaluate the BLEU score of the target phrase *e* in the noisy phrase table.
4. Similarly, for each *e* in (*e*, *f*), identify (*e*, *f**) in the gold phrase table and compute the BLEU score of *f* using *f** as the reference.
5. The sum of the above two BLEU scores is the phrase pair's translation quality score.

5 Phrase Rescoring

Given the bilingual phrase pairs' quality score, there are several ways to use them for statistical machine translation.

5.1 Quality score as a decoder feature

A straightforward way is to use the quality scores as an additional feature in the SMT system, combined with other features (phrase scores, word scores, distortion scores, LM scores etc.) for MT hypotheses scoring. The feature weight can be empirically learned using manual tuning or automatic tuning such as MERT (Och 2003). In this situation, all the phrase pairs and their quality scores are stored in the MT system, which is different from the following approach where incorrect phrase translations are pruned.

5.2 Discriminative phrase rescoring

Another approach is to select good and bad phrase pairs based on their predicted quality scores, then discriminatively rescore the phrase pairs in the baseline phrase library. We sort the phrase pairs based on their quality scores in a decreasing order. The bottom *N* phrase pairs are

considered as incorrect translations and pruned from the phrase library. The top M phrase pairs P_M are considered as good phrases with correct translations. As identifying correct sub-phrase translation requires accurate word alignment within phrase pairs, which is not easy to obtain due to the lack of rich context information within the phrase pair, we only boost the good phrase pairs' super-phrases in the phrase library. Given a phrase pair (e,f) with phrase co-occurrence count $C(e,f)$, the weighted co-occurrence count is defined as:

$$C'(e,f) = C(e,f) \prod_{(e_i,f_i)\in(e,f)} b_i \qquad (9)$$

where (e_i, f_i) is a *good* sub-phrase pair of (e,f) belonging to P_M, with quality score b_i. Note that if (e,f) contains multiple good sub-phrase pairs, its co-occurrence count will be boosted multiple times. Here the boost factor is defined as the product of quality scores of good sub-phrase pairs. Instead of product, one can also use sum, which did not perform as well in our experiments. The weighted co-occurrence count is used to calculate the new phrase translation scores:

$$P'(e \mid f) = \frac{C'(e,f)}{\sum C'(*,f)} \qquad (10)$$

$$P'(f \mid e) = \frac{C'(e,f)}{\sum C'(e,*)} \qquad (11)$$

which replace the original phrase translation scores in the SMT system. In addition to phrase co-occurrence count rescoring, the quality scores can also be used to rescore word translation lexicons by updating word co-occurrence counts accordingly.

6 Experiments

We conducted several experiments to evaluate the proposed phrase rescoring approach. First we evaluate the two regression models' quality score prediction accuracy. Secondly, we apply the predicted phrase scores on machine translation tasks. We will measure the improvement on translation quality as well as the reduction of model size. Our experiments are on English-Chinese translation.

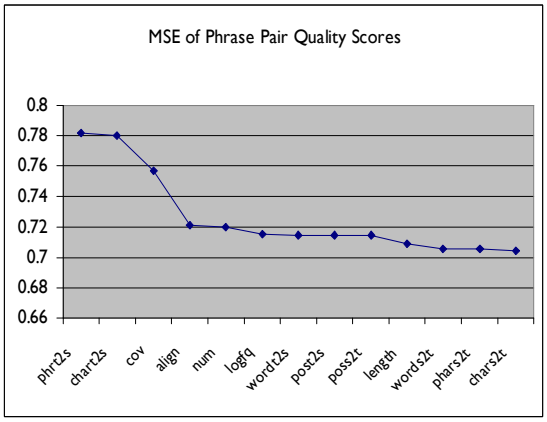

Figure 1. Linear regression model phrase pair prediction MSE curve. Errors are significantly reduced when more features are introduced (phrs2t /phrt2s: phrase source-to-target/target-to-source features; words2t/wordt2s: word-level; chars2t/chart2s: character-level; poss2t/post2s: POS-level; cov: coverage ratio; align: alignment pattern; logfq: log frequency; num: number match; length: length ratio).

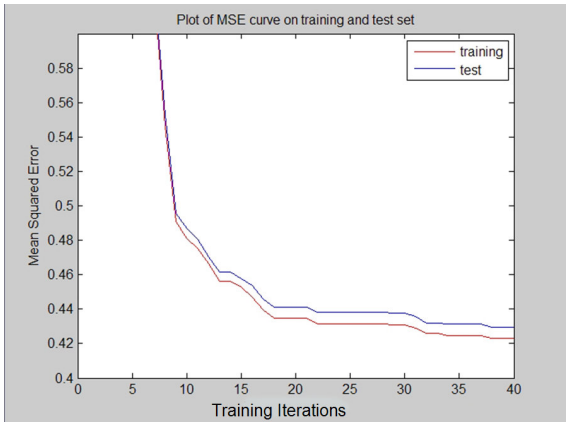

Figure 2. Neural network model phrase pair prediction MSE curve. Errors are significantly reduced with more training iterations.

6.1 Regression model evaluation

We select 10K English-Chinese sentence pairs with both hand alignment and automatic HMM alignment, and extract 106K phrase pairs with true phrase translation quality scores as computed according to formula 8. We choose 53K phrase pairs for regression model training and another 53K phrase pairs for model evaluation. There are 14 parameters to be learned (13 feature weights plus an intercept parameter) for the linear regression model, and 280 weights (13×20

	Linear Regression	Neural Network
Good phrase pairs	and|和|5.52327 amount|金额 数量|4.03006 us|, 美 -|3.91992 her husband|她 丈夫|3.85536 the program|节目 , 一|3.81078 the job|了 这 份 工作|3.77406 shrine|; 靖国神社|3.74336 of course ,|, 当然 , 就 是|3.7174 is only|只 能 是 这|3.69426 visit|访问 只|3.67256 facilities and|设施 , 并 在|3.65402	rights|权利 |6.96817 has become|已 成为 |4.16468 why|为甚么 |3.82629 by armed|受 武装 |3.62988 o|O |3.47795 of drama|在 戏剧 |3.36601 government and|政府 及 |3.27347 introduction|引进 |3.19113 heart disease|心脏 疾病 |3.11829 heads|首脑们 |3.05467 american consumers|美国 消费者 |2.99706
Bad phrase pairs	as well|及 其|1.03234 closed|落下 帷幕|1.01271 she was|梅克尔|0.99011 way|改为 双程|0.955918 of a|出 一 种|0.914717 knowledge|察觉|0.875116 made|出席 "|0.837358 the|保持 联络|0.801142 end|之前|0.769938 held|而 进行 的|0.742588	letter|致函 贵会 |0.39203 , though|尽管 它 |0.37020 levels of|各 级 落实 |0.34892 - board|面板 |0.32826 number of|批 举报 |0.30499 indonesia|苏马尔佐托 |0.27827 xinhua at|$num |0.24433 provinces|安徽 |0.20281 new .|新鲜 之 处 的 , |0.15430 can|的 不同 |0.09502

Table 2. Examples of good and bad phrase pairs based on the linear regression model and neural network's predicted quality scores.

for the input weight matrix plus 20×1 for the output weight vector) for the neural network model. In both cases, the training data size is much more than the parameters size, so there is no data sparseness problem.

After the model parameters are learned from the training data, we apply the regression model to the evaluation data set, then compute the phrase quality score prediction mean squared error (MSE, also known as the average residual sum of squares):

$$MSE = \frac{1}{K} \sum_{k} \left[B_p(e_k, f_k) - B_t(e_k, f_k) \right]^2 \quad (12)$$

where B_p is the predicted quality score of the phrase pair (e_k, f_k), while B_t is the true score calculated based on human translations.

Figure 1 shows the reduction of the regression error in the linear regression model trained with different features. One may find that the MSE is significantly reduced (from 0.78 to 0.70) when additional features are added into the regression model.

Similarly, the neural network's MSE curve is shown in Figure 2. It can be seen that the MSE is significantly reduced with more iterations of training (from the initial error of 1.33 to 0.42 after 40 iterations).

Table 2 shows some phrase pairs with high/low quality scores predicted by the linear regression model and the neural network. One can see that both models assign high scores to good phrase translations and low scores to noisy phrase pairs. Although the values of these scores are beyond the range of [0, 2] as defined in formula 8, this is not a problem for our MT tasks, since they are only used as phrase boosting weights or pruning threshold.

6.2 Machine translation evaluation

We test the above phrase rescoring approach on English-Chinese machine translation. The SMT system is a phrase-based decoder similar to the description in (Tillman 2006), where various features are combined within the log-linear framework. These features include source-to-target phrase translation score based on relative frequency, source-to-target and target-to-source word-to-word translation scores, language model score, distortion model scores and word count. The training data for these features are 10M Chi-

	BLEU	NIST	Phrase Table Size
Baseline	38.67	9.3738	3.65M
LR-mtfeat	39.31	9.5356	3.65M
LR-boost (top30k)	39.36	9.5465	3.65M
LR-prune (tail600k)	39.06	9.4890	3.05M
LR-disc (top30K/tail600K)	39.75	9.6388	3.05M
NN-disc (top30K/tail600K)	39.76	9.6547	3.05M
LR-disc tuning	**39.87**	**9.6594**	**3.05M**
Significance-prune	38.96	9.3953	3.01M
Count-Prune	38.65	9.3549	3.05M

Table 3. Translation quality improvements with rescored phrase tables. Best result (1.2 BLEU gain) is obtained with discriminative rescoring by boosting top 30K phrase pairs and pruning bottom 600K phrase pairs, with some weight tuning.

nese-English sentence pairs, mostly newswire and UN corpora released by LDC. The parallel sentences have word alignment automatically generated with HMM and MaxEnt word aligner. Bilingual phrase translations are extracted from these word-aligned parallel corpora. Due to the noise in the bilingual sentence pairs and automatic word alignment errors, the phrase translation library contains many incorrect phrase translations, which lead to inaccurate translations, as seen in Figure 3.

Our evaluation data is NIST MT08 English-Chinese evaluation testset, which includes 1859 sentences from 129 news documents. The automatic metrics are BLEU and NIST scores, as used in the NIST 2008 English-Chinese MT evaluation. Note that as there is no whitespace as Chinese word boundary, the Chinese translations are segmented into characters before scoring in order to reduce the variance and errors caused by automatic word segmentation, which is also done in the NIST MT evaluation.

Table 3 shows the automatic MT scores using the baseline phrase table and rescored phrase tables. When the phrase quality scores from the linear regression model are used as a separate feature in the SMT system (***LR-mtfeat*** as described in section 5.1), the improvement is 0.7 BLEU points (0.16 in terms of NIST scores). By boosting the good phrase pairs (top 30K[2] phrase pairs, ***LR-boost***) from linear regression model, the MT quality is improved by 0.7 BLEU points over the baseline system. Pruning the bad phrase pairs (tail 600K phrase pairs) without using the quality scores as features (***LR-prune***) also improves the MT by 0.4 BLEU points. Combining ***LR-boost*** and ***LR_prune***, a discriminatively rescored phrase table (***LR-disc***) improved the BLEU score by 1.1 BLEU points, and reduce the phrase table size by 16% (from 3.6M to 3.0M phrase pairs). Manually tuning the boosting weights of good phrase pairs leads to additional improvement. Discriminative rescoring using the neural net work scores (***NN-disc***) produced similar improvement**.**

We also experiment with phrase table pruning using Fisher significant test, as proposed in (Johnson et. al. 2007). We tuned the pruning threshold for the best result. It shows that the significance pruning improves over the baseline by 0.3 BLEU pts with 17.5% reduction in phrase table, but is not as good as our proposed phrase rescoring method. In addition, we also show the MT result using a count pruning phrase table (Count-Prune) where 600K phrase translation pairs are pruned based on their co-occurrence counts. The MT performance of such phrase table pruning is slightly worse than the baseline MT system, and significantly worse than the result using the proposed rescored phrase table.

When comparing the linear regression and neural network models, we find rescoring with both models lead to similar MT improvements, even though the neural network model has much fewer regression errors (0.44 vs. 0.7 in terms of MSE). This is due to the rich parameter space of the neural network.

Overall, the discriminative phrase rescoring improves the SMT quality by 1.2 BLEU points and reduces the phrase table size by 16%. With statistical significance test (Zhang and Vogel 2004), all the improvements are statistically significant with p-value < 0.0001.

Figure 3 presents some English sentences, with phrase translation pairs selected in the final translations (the top one is from the baseline MT system and the bottom one is from the LR-disc system).

[2] These thresholds are empirically chosen.

Src	Indonesian bird flu victim contracted virus indirectly:
Baseline	<indonesian bird flu|印尼 禽流感> <virus|病毒> **<victim contracted|感染者>** <indirectly :|间接 :>
PhrResco	<indonesian bird flu|印尼 禽流感> **<victim|受害者> <contracted|感染>** <virus|病毒> <indirectly :|间接 :>
Src	The director of Palestinian human rights group Al-Dhamir, Khalil Abu Shammaleh, said he was also opposed to the move.
Baseline	<the director of|署长 的> <palestinian|巴勒斯坦> <human rights group|人权 团体> **<al -|" 基地 " 组织>** <,|,> **<abu|Abu>** <khalil|Khalil> <, said he was|表示 , 他> <also opposed to|也 反对> <the move .|这 项 行动 。>
PhrResco	<the director of|署长 的> <palestinian|巴勒斯坦> <human rights group|人权 团体> **<al -|al ->** <, khalil|, khalil> **<abu|阿布>** <, said he was|说 , 他> <also opposed to|也 反对> <the move .|这 项 行动 。>
Src	A young female tourist and two of her Kashmiri friends were among the victims.
Baseline	<a young female|有 一 名 年轻 女子> **<tourist and|旅游 和>** <$num of her|她 的 $num 个> <kashmiri|克什米尔> <friends were|网友> <among the|之间 的> <victims .|受害者 。>
PhrResco	<a young|一 个 年轻 的> <female|女性> **<tourist and|游客 和>** <$num of her|她 的 $num 个> <kashmiri|克什米尔> <friends were|朋友> <among the|之间 的> <victims .|受害者 。>

Figure 3. Examples of English sentences and their translation, with phrase pairs from baseline system and phrase rescored system. Highlighted text are initial phrase translation errors which are corrected in the PhrResco translations.

We find that incorrect phrase translations in the baseline system (as highlighted with blue bold font) are corrected and better translation results are obtained.

7 Conclusion

We introduced a discriminative phrase rescoring approach, which combined rich features with linear regression and neural network to predict phrase pair translation qualities. Based on these quality scores, we boost good phrase translations while pruning bad phrase translations. This led to statistically significant improvement (1.2 BLEU points) in MT and reduced phrase table size by 16%.

For the future work, we would like to explore other models for quality score prediction, such as SVM. We will want to try other approaches to utilize the phrase pair quality scores, in addition to rescoring the co-occurrence frequency. Finally, we will test this approach in other domain applications and language pairs.

References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*, Computational Linguistics, v.19 n.2, June 1993.

Arthur Earl Bryson, Yu-Chi Ho. 1969. *Applied Optimal Control: Optimization, Estimation, and Control*. Blaisdell Publishing Company. p481.

David Chiang. 2005. *A Hierarchical Phrase-based Model for Statistical Machine Translation*. 2005. In Proc. of ACL, pp. 263–270.

Yonggang Deng, Jia Xu, and Yuqing Gao. 2008. *Phrase Table Training for Precision and Recall: What Makes a Good Phrase and a Good Phrase Pair*? In Proc. of ACL/HLT, pp. 81-88.

Michel Galley, Mark Hopkins, Kevin Knight, Daniel Marcu. 2004. *What's in a Translation Rule?* In Proc. of NAACL 2004, pp. 273-280.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. *Improving Translation Quality by Discarding Most of the Phrase Table.* In Proc. of EMNLP-CoNLL, pp. 967-975.

Philipp Koehn, Franz Josef Och, Daniel Marcu. 2003. *Statistical Phrase-based Translation*, In Proc. of NAACL, pp. 48-54.

Franz Josef Och and Hermann Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, v.29 n.1, pp.19-51, March 2003

Franz Josef Och. 2003. *Minimum Error Rate Training in Statistical Machine Translation*, In Proc. of ACL, 2003, pp. 160-167.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*, In Proc. of ACL, pp. 311-318.

Christoph Tillmann. 2006. *Efficient Dynamic Programming Search Algorithms for Phrase-based SMT*. In Proc. of the Workshop CHPSLP at HLT'06.

Kenji Yamada and Kevin Knight. 2001. *A Syntax-based Statistical Translation Model*, In Proc. of ACL, pp.523-530.

Mei Yang and Jing Zheng. 2009. *Toward Smaller, Faster, and Better Hierarchical Phrase-based SMT*. In Proc. of ACL-IJCNLP, pp. 237-240.

Ying Zhang and Stephan Vogel. 2004. *Measuring Confidence Intervals for the Machine Translation Evaluation Metrics*, In Proc. TMI, pp. 4-6.

Bing Zhao, Stephan Vogel, and Alex Waibel. 2004. *Phrase Pair Rescoring with Term Weighting for Statistical Machine Translation*. In Proc. of EMNLP, pp. 206-213.

Andreas Zollmann and Ashish Venugopal. 2006. *Syntax Augmented Machine Translation via Chart Parsing*. In Proc. of NAACL 2006- Workshop on statistical machine translation.