Effective Use of Linguistic and Contextual Information for Statistical Machine Translation

Libin Shen and **Jinxi Xu** and **Bing Zhang** and
Spyros Matsoukas and **Ralph Weischedel**
BBN Technologies
Cambridge, MA 02138, USA
{lshen,jxu,bzhang,smatsouk,weisched}@bbn.com

Abstract

Current methods of using lexical features in machine translation have difficulty in scaling up to realistic MT tasks due to a prohibitively large number of parameters involved. In this paper, we propose methods of using new linguistic and contextual features that do not suffer from this problem and apply them in a state-of-the-art hierarchical MT system. The features used in this work are non-terminal labels, non-terminal length distribution, source string context and source dependency LM scores. The effectiveness of our techniques is demonstrated by significant improvements over a strong baseline. On Arabic-to-English translation, improvements in lower-cased BLEU are 2.0 on NIST MT06 and 1.7 on MT08 newswire data on decoding output. On Chinese-to-English translation, the improvements are 1.0 on MT06 and 0.8 on MT08 newswire data.

1 Introduction

Linguistic and context features, especially sparse lexical features, have been widely used in recent machine translation (MT) research. Unfortunately, existing methods of using such features are not ideal for large-scale, practical translation tasks.

In this paper, we will propose several probabilistic models to effectively exploit linguistic and contextual information for MT decoding, and these new features do not suffer from the scalability problem. Our new models are tested on NIST MT06 and MT08 data, and they provide significant improvement over a strong baseline system.

1.1 Previous Work

The ideas of using labels, length preference and source side context in MT decoding were explored previously. Broadly speaking, two approaches were commonly used in existing work.

One is to use a stochastic gradient descent (SGD) or Perceptron like online learning algorithm to optimize the weights of these features directly for MT (Shen et al., 2004; Liang et al., 2006; Tillmann and Zhang, 2006). This method is very attractive, since it opens the door to rich lexical features. However, in order to robustly optimize the feature weights, one has to use a substantially large development set, which results in significantly slower tuning. Alternatively, one needs to carefully select a development set that simulates the test set to reduce the risk of over-fitting, which however is not always realistic for practical use.

A remedy is to aggressively limit the feature space, e.g. to syntactic labels or a small fraction of the bi-lingual features available, as in (Chiang et al., 2008; Chiang et al., 2009), but that reduces the benefit of lexical features. A possible generic solution is to cluster the lexical features in some way. However, how to make it work on such a large space of bi-lingual features is still an open question.

The other approach is to estimate a single score or likelihood of a translation with rich features, for example, with the maximum entropy (MaxEnt) method as in (Carpuat and Wu, 2007; Ittycheriah and Roukos, 2007; He et al., 2008). This method avoids the over-fitting problem, at the expense of losing the benefit of discriminative training of rich features directly for MT. However, the feature space problem still exists in these published models.

He et al. (2008) extended the WSD-like approached proposed in (Carpuat and Wu, 2007) to hierarchical decoders. In (He et al., 2008), lexical

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 72–80,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

features were limited on each single side due to the feature space problem. In order to further reduce the complexity of MaxEnt training, they "trained a MaxEnt model for each ambiguous hierarchical LHS" (left-hand side or source side) of translation rules. Different target sides were treated as possible labels. Therefore, the sample sets of each individual MaxEnt model were very small, while the number of features could easily exceed the number of samples. Furthermore, optimizing individual MaxEnt models in this way does not lead to global maximum. In addition, MaxEnt models trained on small sets are unstable.

The MaxEnt model in (Ittycheriah and Roukos, 2007) was optimized globally, so that it could better employ the distribution of the training data. However, one has to filter the training data according to the test data to get competitive performance with this model [1]. In addition, the filtering method causes some practical issues. First, such methods are not suitable for real MT tasks, especially for applications with streamed input, since the model has to be retrained with each new input sentence or document and training is slow. Furthermore, the model is ill-posed. The translation of a source sentence depends on other source sentences in the same batch with which the MaxEnt model is trained. If we add one more sentence to the batch, translations of other sentences may become different due to the change of the MaxEnt model.

To sum up, the existing models of employing rich bi-lingual lexical information in MT are imperfect. Many of them are not ideal for practical translation tasks.

1.2 Our Approach

As for our approach, we mainly use simple probabilistic models, i.e. Gaussian and n-gram models, which are more robust and suitable for large-scale training of real data, as manifested in state-of-the-art systems of speech recognition. The unique contribution of our work is to design effective and efficient statistical models to capture useful linguistic and context information for MT decoding. Feature functions defined in this way are robust and ideal for practical translation tasks.

1.2.1 Features

In this paper, we will introduce four new linguistic and contextual feature functions. Here, we first provide a high-level description of these features. Details of the features are discussed in Section 2.

The first feature is based on non-terminal labels, i.e. POS tags of the head words of target non-terminals in transfer rules. This feature reduces the ambiguity of translation rules. The other benefit is that POS tags help to weed out bad target side tree structures, as an enhancement to the target dependency language model.

The second feature is based on the length distribution of non-terminals. In English as well as in other languages, the same deep structure can be represented in different syntactic structures depending on the complexity of its constituents. We model such preferences by associating each non-terminal of a transfer rule with a probability distribution over its length. Similar ideas were explored in (He et al., 2008). However their length features only provided insignificant improvement of 0.1 BLEU point. A crucial difference of our approach is how the length preference is modeled. We approximate the length distribution of non-terminals with a smoothed Gaussian, which is more robust and gives rise to much larger improvement consistently.

The third feature utilizes source side context information, i.e. the neighboring words of an input span, to influence the selection of the target translation for a span. While the use of context information has been explored in MT, e.g. (Carpuat and Wu, 2007) and (He et al., 2008), the specific technique we used by means of a context language model is rather different. Our model is trained on the whole training data, and it is not limited by the constraint of MaxEnt training.

The fourth feature exploits structural information on the source side. Specifically, the decoder simultaneously generates both the source and target side dependency trees, and employs two dependency LMs, one for the source and the other for the target, for scoring translation hypotheses. Our intuition is that the likelihood of source structures provides another piece of evidence about the plausibility of a translation hypothesis and as such would help weed out bad ones.

[1] According to footnote 2 of (Ittycheriah and Roukos, 2007), test set adaptation by test set sampling of the training corpus showed an advantage of more than 2 BLEU points over a general system trained on all data.

1.2.2 Baseline System and Experimental Setup

We take BBN's *HierDec*, a string-to-dependency decoder as described in (Shen et al., 2008), as our baseline for the following two reasons:

- It provides a strong baseline, which ensures the validity of the improvement we would obtain. The baseline model used in this paper showed state-of-the-art performance at NIST 2008 MT evaluation.
- The baseline algorithm can be easily extended to incorporate the features proposed in this paper. The use of source dependency structures is a natural extension of the string-to-tree model to a tree-to-tree model.

To ensure the generality of our results, we tested the features on two rather different language pairs, Arabic-to-English and Chinese-to-English, using two metrics, IBM BLEU (Papineni et al., 2001) and TER (Snover et al., 2006). Our experiments show that each of the first three features: non-terminal labels, length distribution and source side context, improves MT performance. Surprisingly, the source dependency feature does not produce an improvement.

2 Linguistic and Context Features

2.1 Non-terminal Labels

In the original string-to-dependency model (Shen et al., 2008), a translation rule is composed of a string of words and non-terminals on the source side and a well-formed dependency structure on the target side. A well-formed dependency structure could be either a single-rooted dependency tree or a set of sibling trees. As in the Hiero system (Chiang, 2007), there is only one non-terminal X in the string-to-dependency model. Any sub dependency structure can be used to replace a non-terminal in a rule.

For example, we have a source sentence in Chinese as follows.

- jiantao zhuyao baohan liang fangmian

The literal translation for individual words is

- 'review' 'mainly' 'to consist of' 'two' 'part'

The reference translation is

- the review mainly consists of two parts

A single source word can be translated into many English words. For example, *jiantao* can be translated into *a review, the review, reviews, the reviews, reviewing, reviewed*, etc. Suppose we have source-string-to-target-dependency translation rules as shown in Figure 1. Since there is no constraint on substitution, any translation for *jiantao* could replace the X-1 slot.

One way to alleviate this problem is to limit the search space by using a label system. We could assign a label to each non-terminal on the target side of the rules. Furthermore, we could assign a label to the whole target dependency structure, as shown in Figure 2. In decoding, each target dependency sub-structure would be associated with a label. Whenever substitution happens, we would check whether the label of the sub-structure and the label of the slot are the same. Substitutions with unmatched labels would be prohibited.

In practice, we use a soft constraint by penalizing substitutions with unmatched labels. We introduce a new feature: the number of times substitutions with unmatched labels appear in the derivation of a translation hypothesis.

Obviously, to implement this feature we need to associate a label with each non-terminal in the target side of a translation rule. The labels are generated during rule extraction. When we create a rule from a training example, we replace a subtree or dependency structure with a non-terminal and associate it with the POS tag of the head word if the non-terminal corresponds to a single-rooted tree on the target side. Otherwise, it is assigned the generic label X. (In decoding, all substitutions of X are considered unmatched ones and incur a penalty.)

2.2 Length Distribution

In English, the length of a phrase may determine the syntactic structure of a sentence. For example, possessive relations can be represented either as "A's B" or "B of A". The former is preferred if A is a short phrase (e.g. "the boy's mother") while the latter is preferred if A is a complex structure (e.g. "the mother of the boy who is sick").

Our solution is to build a model of length preference for each non-terminal in each translation rule. To address data sparseness, we assume the length distribution of each non-terminal in a transfer rule is a Gaussian, whose mean and variance can be estimated from the training data. In rule extrac-

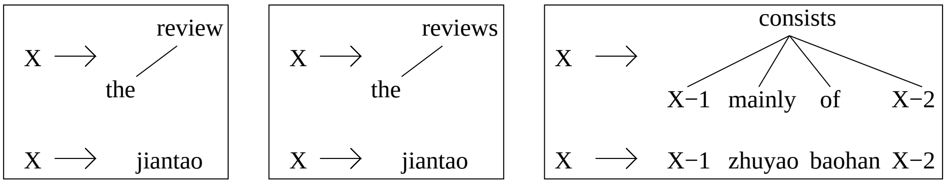

Figure 1: Translation rules with one label X

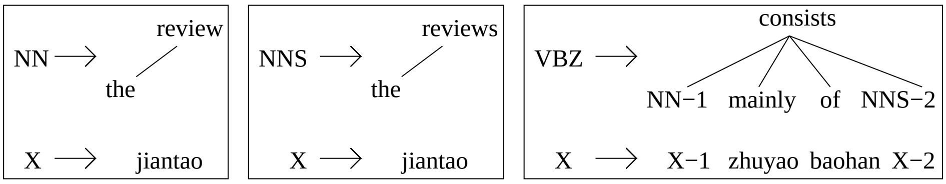
NN ⟶ the review
X ⟶ jiantao
NNS ⟶ the reviews
X ⟶ jiantao
VBZ ⟶ consists NN−1 mainly of NNS−2
X ⟶ X−1 zhuyao baohan X−2

Figure 2: Translation rules with multiple labels

tion, each time a translation rule is generated from a training example, we can record the length of the source span corresponding to a non-terminal. In the end, we have a frequency histogram for each non-terminal in each translation rule. From the histogram, a Gaussian distribution can be easily computed.

In practice, we do not need to collect the frequency histogram. Since all we need to know are the mean and the variance, it is sufficient to collect the sum of the length and the sum of squared length.

Let r be a translation rule that occurs N_r times in training. Let x be a specific non-terminal in that rule. Let $l(r, x, i)$ denote the length of the source span corresponding to non-terminal x in the i-th occurrence of rule r in training. Then, we can compute the following quantities.

$$m_{r,x} = \frac{1}{N_r}\sum_{i=1}^{N_r} l(r,x,i) \tag{1}$$

$$s_{r,x} = \frac{1}{N_r}\sum_{i=1}^{N_r} l(r,x,i)^2, \tag{2}$$

which can be subsequently used to estimate the mean $\mu_{r,x}$ and variance $\sigma^2_{r,x}$ of x's length distribution in rule r as follows.

$$\mu_{r,x} = m_{r,x} \tag{3}$$

$$\sigma^2_{r,x} = s_{r,x} - m^2_{r,x} \tag{4}$$

Since many of the translation rules have few occurrences in training, smoothing of the above estimates is necessary. A common smoothing method is based on maximum a posteriori (MAP) estimation as in (Gauvain and Lee, 1994).

$$\begin{aligned}\hat{m}_{r,x} &= \frac{N_r}{N_r+\tau}m_{r,x} + \frac{\tau}{N_r+\tau}\tilde{m}_{r,x}\\ \hat{s}_{r,x} &= \frac{N_r}{N_r+\tau}s_{r,x} + \frac{\tau}{N_r+\tau}\tilde{s}_{r,x},\end{aligned}$$

where ˆ stands for an MAP distribution and ˜ represents a prior distribution. $\tilde{m}_{r,x}$ and $\tilde{s}_{r,x}$ can be obtained from a prior Gaussian distribution $\mathcal{N}(\tilde{\mu}_{r,x}, \tilde{\sigma}_{r,x})$ via equations (3) and (4), and τ is a weight of smoothing.

There are many ways to approximate the prior distribution. For example, we can have one prior for all the non-terminals or one for individual non-terminal type. In practice, we assume $\tilde{\mu}_{r,x} = \mu_{r,x}$, and approximate $\tilde{\sigma}_{r,x}$ as $(\sigma^2_{r,x} + s_{r,x})^{\frac{1}{2}}$.

In this way, we do not change the mean, but relax the variance with $s_{r,x}$. We tried different smoothing methods, but the performance did not change much, therefore we kept this simplest setup. We also tried the Poisson distribution, and the performance is similar to Gaussian distribution, which is about 0.1 point lower in BLEU.

When a rule r is applied during decoding, we compute a penalty for each non-terminal x in r according to

$$P(l \mid r, x) = \frac{1}{\sigma_{r,x}\sqrt{2\pi}} e^{-\frac{(l-\mu_{r,x})^2}{2\sigma^2_{r,x}}},$$

where l is length of source span corresponding to x.

Our method to address the problem of length bias in rule selection is very different from the maximum entropy method used in existing studies, e.g. (He et al., 2008).

2.3 Context Language Model

In the baseline string-to-dependency system, the probability a translation rule is selected in decoding does not depend on the sentence context. In reality, translation is highly context dependent. To address this defect, we introduce a new feature, called *context language model*. The motivation of this feature is to exploit surrounding words to influence the selection of the desired transfer rule for a given input span.

To illustrate the problem, we use the same example mentioned in Section 2.1. Suppose the source span for rule selection is *zhuyao baohan*, whose literal translation is *mainly* and *to consist of*. There are many candidate translations for this phrase, for example, *mainly consist of, mainly consists of, mainly including, mainly includes*, etc. The surrounding words can help to decide which translation is more appropriate for *zhuyao baohan*. We compare the following two context-based probabilities:

- $P($ jiantao $|$ mainly consist $)$
- $P($ jiantao $|$ mainly consists $)$

Here, *jiantao* is the source word preceding the source span *zhuyao baohan*.

In the training data, *jiantao* is usually translated into *the review*, third-person singular, then the probability $P($ jiantao $|$ mainly consists $)$ will be higher than $P($ jiantao $|$ mainly consist $)$, since we have seen more context events like the former in the training data.

Now we introduce context LM formally. Let the source words be $f_1 f_2..f_i..f_j..f_n$. Suppose source sub-string $f_i..f_j$ is translated into $e_p..e_q$. We can define tri-gram probabilities on the left and right sides of the source span:

- left : $P_L(f_{i-1}|e_p, e_{p+1})$
- right : $P_R(f_{j+1}|e_q, e_{q-1})$

In our implementation, the left and right context LMs are estimated from the training data as part of the rule extraction procedure. When we exact a rule, we collect two 3-gram events, one for the left side and the other for the right side.

In decoding, whenever a partial hypothesis is generated, we calculate the context LM scores based on the leftmost two words and the rightmost two words of the hypothesis as well as the source context. The product of the left and right context LM scores is used as a new feature in the scoring function.

Please note that our approach is very different from other approaches to context dependent rule selection such as (Ittycheriah and Roukos, 2007) and (He et al., 2008). Instead of using a large number of fine grained features with weights optimized using the maximum entropy method, we treat context dependency as an ngram LM problem, and it is smoothed with Witten-Bell discounting. The estimation of the context LMs is very efficient and robust.

The benefit is two fold. The estimation of the context LMs is very efficient. It adds only one new weight to the scoring function.

2.4 Source Dependency Language Model

The context LM proposed in the previous section only employs source words immediately before and after the current source span in decoding. To exploit more source context, we use a source side dependency language model as another feature. The motivation is to take advantage of the long distance dependency relations between source words in scoring a translation theory.

We extended string-to-dependency rules in the baseline system to dependency-to-dependency rules. In each dependency-to-dependency rule, we keep record of the source string as well as the source dependency structure. Figure 3 shows examples of dependency-to-dependency rules.

We extended the string-to-dependency decoding algorithm in the baseline to accommodate dependency-to-dependency theories. In decoding, we build both the source and the target dependency structures simultaneously in chart parsing over the source string. Thus, we can compute the source dependency LM score in the same way we compute the target side score, using a procedure described in (Shen et al., 2008).

We introduce two new features for the source side dependency LM as follows, in a way similar to the target side.

- Source dependency LM score
- Discount on ill-formed source dependency structures

The source dependency LM is trained on the source side of the bi-lingual training data with Witten-Bell smoothing. The source dependency LM score represents the likelihood of the source

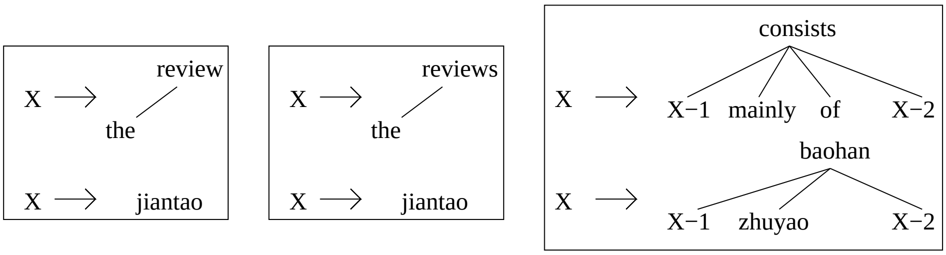

Figure 3: Dependency-to-dependency translation rules

dependency tree generated by the decoder. The source dependency tree with the highest score is the one that is most likely to be generated by the dependency model that created the source side of the training data.

Source dependency trees are composed of fragments embedded in the translation rules. Therefore, a source dependency LM score can be viewed as a measure whether the translation rules are put together in a way similar to the training data. Therefore, a source dependency LM score serves as a feature to represent structural context information that is capable of modeling long-distance relations.

However, unlike source context LMs, the structural context information is used only when two partial dependency structures are combined, while source context LMs work as a look-ahead feature.

3 Experiments

We designed our experiments to show the impact of each feature separately as well as their cumulative impact:

- BASE: baseline string-to-dependency system
- SLM: baseline + source dependency LM
- CLM: baseline + context LM
- LEN: baseline + length distribution
- LBL: baseline + syntactic labels
- LBL+LEN: baseline + syntactic labels + length distribution
- LBL+LEN+CLM: baseline + syntactic labels + length distribution + context LM

All the models were optimized on lower-cased IBM BLEU with Powell's method (Powell, 1964; Brent, 1973) on n-best translations (Ostendorf et al., 1991), but evaluated on both IBM BLEU and TER. The motivation is to detect if an improvement is artificial, i.e., specific to the tuning metric. For both Arabic-to-English and Chinese-to-English MT, we tuned on NIST MT02-05 and tested on MT06 and MT08 newswire sets.

The training data are different from what was usd at MT06 or MT08. Our Arabic-to-English data contain 29M Arabic words and 38M English words from 11 corpora: LDC2004T17, LDC2004T18, LDC2005E46, LDC2006E25, LDC2006G05, LDC2005E85, LDC2006E36, LDC2006E82, LDC2006E95, Sakhr-A2E and Sakhr-E2A. The Chinese-to-English data contain 107M Chinese words and 132M English words from eight corpora: LDC2002E18, LDC2005T06, LDC2005T10, LDC2006E26, LDC2006G05, LDC2002L27, LDC2005T34 and LDC2003E07. They are available under the DARPA GALE program. Traditional 3-gram and 5-gram string LMs were trained on the English side of the parallel data plus the English Gigaword corpus V3.0 in a way described in (Bulyko et al., 2007).

The target dependency LMs were trained on the English side of the parallel training data. For that purpose, we parsed the English side of the parallel data. Two separate models were trained: one for Arabic from the Arabic training data and the other for Chinese from the Chinese training data.

To compute the source dependency LM for Chinese-to-English MT, we parsed the Chinese side of the Chinese-to-English parallel data. Due to the lack of a good Arabic parser compatible with the Sakhr tokenization that we used on the source side, we did not test the source dependency LM for Arabic-to-English MT.

When extracting rules with source dependency structures, we applied the same *well-formedness* constraint on the source side as we did on the target side, using a procedure described by (Shen et al., 2008). Some candidate rules were thrown away due to the source side constraint. On the

	MT06				**MT08**			
Model	**BLEU**		**TER**		**BLEU**		**TER**	
	lower	**mixed**	**lower**	**mixed**	**lower**	**mixed**	**lower**	**mixed**
Decoding (3-gram LM)								
BASE	48.75	46.74	43.43	45.79	49.58	47.46	42.80	45.08
CLM	49.44	47.36	42.96	45.22	49.73	47.53	42.64	44.92
LEN	49.37	47.28	43.01	45.35	50.29	48.19	42.32	44.45
LBL	49.33	47.07	43.09	45.53	50.46	48.19	42.27	44.57
LBL+LEN	49.91	47.70	42.59	45.17	51.10	48.85	41.88	44.16
LBL+LEN+CLM	50.75	48.51	42.13	44.50	51.24	49.10	41.63	43.80
Rescoring (5-gram LM)								
BASE	51.24	49.23	42.08	44.42	51.23	49.11	42.01	44.15
CLM	51.57	49.54	41.74	43.88	51.44	49.37	41.63	43.74
LEN	52.05	50.01	41.50	43.72	51.88	49.89	41.51	43.47
LBL	51.80	49.69	41.54	43.76	51.93	49.86	41.27	43.33
LBL+LEN	51.90	49.76	41.41	43.70	52.42	50.29	40.93	43.00
LBL+LEN+CLM	52.61	50.51	40.77	43.03	52.60	50.56	40.69	42.81

Table 1: BLEU and TER percentage scores on MT06 and MT08 Arabic-to-English newswire sets.

other hand, one string-to-dependency rule may split into several dependency-to-dependency rules due to different source dependency structures. The size of the dependency-to-dependency rule set is slightly smaller than the size of the string-to-dependency rule set.

Tables 1 and 2 show the BLEU and TER percentage scores on MT06 and MT08 for Arabic-to-English and Chinese-to-English translation respectively. The context LM feature, the length feature and the syntax label feature all produce a small improvement for most of the conditions. When we combined the three features, we observed significant improvements over the baseline. For Arabic-to-English MT, the LBL+LEN+CLM system improved lower-cased BLEU by 2.0 on MT06 and 1.7 on MT08 on decoding output. For Chinese-to-English MT, the improvements in lower-cased BLEU were 1.0 on MT06 and 0.8 on MT08. After re-scoring, the improvements became smaller, but still noticeable, ranging from 0.7 to 1.4. TER scores were also improved noticeably for all conditions, suggesting there was no metric specific over-tuning.

Surprisingly, source dependency LM did not provide any improvement over the baseline. There are two possible reasons for this. One is that the source and target parse trees were generated by two stand-alone parsers, which may cause incompatible structures on the source and target sides. By applying the *well-formed* constraints on both sides, a lot of useful transfer rules are discarded. A bi-lingual parser, trained on parallel treebanks recently made available to the NLP community, may overcome this problem. The other is that the search space of dependency-to-dependency decoding is much larger, since we need to add source dependency information into the chart parsing states. We will explore techniques to address these problems in the future.

4 Discussion

Linguistic information has been widely used in SMT. For example, in (Wang et al., 2007), syntactic structures were employed to reorder the source language as a pre-processing step for phrase-based decoding. In (Koehn and Hoang, 2007), shallow syntactic analysis such as POS tagging and morphological analysis were incorporated in a phrasal decoder.

In ISI's syntax-based system (Galley et al., 2006) and CMU's Hiero extension (Venugopal et al., 2007), non-terminals in translation rules have labels, which must be respected by substitutions during decoding. In (Post and Gildea, 2008; Shen et al., 2008), target trees were employed to improve the scoring of translation theories. Marton and Resnik (2008) introduced features defined on constituent labels to improve the Hiero system (Chiang, 2005). However, due to the limitation of MER training, only part of the feature space could used in the system. This problem was fixed by

	MT06				**MT08**			
Model	**BLEU**		**TER**		**BLEU**		**TER**	
	lower	**mixed**	**lower**	**mixed**	**lower**	**mixed**	**lower**	**mixed**
Decoding (3-gram LM)								
BASE	37.44	35.62	54.64	56.47	33.05	31.26	56.79	58.69
SLM	37.30	35.48	54.24	55.90	33.03	31.00	56.59	58.46
CLM	37.66	35.81	53.45	55.19	32.97	31.01	55.99	57.77
LEN	38.09	36.26	53.98	55.81	33.23	31.34	56.51	58.41
LBL	38.37	36.53	54.14	55.99	33.25	31.34	56.60	58.49
LBL+LEN	38.36	36.59	53.95	55.60	33.72	31.83	56.79	58.65
LBL+LEN+CLM	38.41	36.57	53.83	55.70	33.83	31.79	56.55	58.51
Rescoring (5-gram LM)								
BASE	38.91	37.04	53.65	55.45	34.34	32.32	55.60	57.60
SLM	38.27	36.38	53.64	55.29	34.25	32.28	55.35	57.21
CLM	38.79	36.88	53.09	54.80	35.01	32.98	55.39	57.28
LEN	39.22	37.30	53.34	55.06	34.65	32.70	55.61	57.51
LBL	39.11	37.30	53.61	55.29	35.02	33.00	55.39	57.48
LBL+LEN	38.91	37.17	53.56	55.27	35.03	33.08	55.47	57.46
LBL+LEN+CLM	39.58	37.62	53.21	54.94	35.72	33.63	54.88	56.98

Table 2: BLEU and TER percentage scores on MT06 and MT08 Chinese-to-English newswire sets.

Chiang et al. (2008), which used an online learning method (Crammer and Singer, 2003) to handle a large set of features.

Most SMT systems assume that translation rules can be applied without paying attention to the sentence context. A few studies (Carpuat and Wu, 2007; Ittycheriah and Roukos, 2007; He et al., 2008; Hasan et al., 2008) addressed this defect by selecting the appropriate translation rules for an input span based on its context in the input sentence. The direct translation model in (Ittycheriah and Roukos, 2007) employed syntactic (POS tags) and context information (neighboring words) within a maximum entropy model to predict the correct transfer rules. A similar technique was applied by He et al. (2008) to improve the Hiero system.

Our model differs from previous work on the way in which linguistic and contextual information is used.

5 Conclusions and Future Work

In this paper, we proposed four new linguistic and contextual features for hierarchical decoding. The use of non-terminal labels, length distribution and context LM features gave rise to significant improvement on Arabic-to-English and Chinese-to-English translation on NIST MT06 and MT08 newswire data over a state-of-the-art string-to-dependency baseline. Unlike previous work, we employed robust probabilistic models to capture useful linguistic and contextual information. Our methods are more suitable for practical translation tasks.

In future, we will continue this work in two directions. We will employ a Gaussian model to unify various linguistic and contextual features. We will also improve the dependency-to-dependency method with a better bi-lingual parser.

Acknowledgments

This work was supported by DARPA/IPTO Contract No. HR0011-06-C-0022 under the GALE program.

References

R. P. Brent. 1973. *Algorithms for Minimization Without Derivatives*. Prentice-Hall.

I. Bulyko, S. Matsoukas, R. Schwartz, L. Nguyen, and J. Makhoul. 2007. Language model adaptation in machine translation from speech. In *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

M. Carpuat and D. Wu. 2007. Context-dependent phrasal translation lexicons for statistical machine translation. In *Proceedings of Machine Translation Summit XI*.

D. Chiang, Y. Marton, and P. Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of the 2008 Conference of Empirical Methods in Natural Language Processing*.

D. Chiang, K. Knight, and W. Wang. 2009. 11,001 new features for statistical machine translation. In *Proceedings of the 2009 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL)*.

D. Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

K. Crammer and Y. Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer. 2006. Scalable inference and training of context-rich syntactic models. In *COLING-ACL '06: Proceedings of 44th Annual Meeting of the Association for Computational Linguistics and 21st Int. Conf. on Computational Linguistics*.

J.-L. Gauvain and Chin-Hui Lee. 1994. Maximum a posteriori estimation for multivariate gaussian mixtureobservations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2).

S. Hasan, J. Ganitkevitch, H. Ney, and J. Andrés-Ferrer. 2008. Triplet lexicon models for statistical machine translation. In *Proceedings of the 2008 Conference of Empirical Methods in Natural Language Processing*.

Z. He, Q. Liu, and S. Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *Proceedings of COLING '08: The 22nd Int. Conf. on Computational Linguistics*.

A. Ittycheriah and S. Roukos. 2007. Direct translation model 2. In *Proceedings of the 2007 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

P. Koehn and H. Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Conference of Empirical Methods in Natural Language Processing*.

P. Liang, A. Bouchard-Côté, D. Klein, and B. Taskar. 2006. An end-to-end discriminative approach to machine translation. In *COLING-ACL '06: Proceedings of 44th Annual Meeting of the Association for Computational Linguistics and 21st Int. Conf. on Computational Linguistics*.

Y. Marton and P. Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*.

M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek. 1991. Integration of diverse recognition methodologies through reevaluation of nbest sentence hypotheses. In *Proceedings of the DARPA Workshop on Speech and Natural Language*.

K. Papineni, S. Roukos, and T. Ward. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report, RC22176.

M. Post and D. Gildea. 2008. Parsers as language models for statistical machine translation. In *The Eighth Conference of the Association for Machine Translation in the Americas*.

M. J. D. Powell. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2).

L. Shen, A. Sarkar, and F. J. Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

L. Shen, J. Xu, and R. Weischedel. 2008. A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

C. Tillmann and T. Zhang. 2006. A discriminative global training algorithm for statistical mt. In *COLING-ACL '06: Proceedings of 44th Annual Meeting of the Association for Computational Linguistics and 21st Int. Conf. on Computational Linguistics*.

A. Venugopal, A. Zollmann, and S. Vogel. 2007. An efficient two-pass approach to synchronous-cfg driven statistical mt. In *Proceedings of the 2007 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

C. Wang, M. Collins, and P. Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Conference of Empirical Methods in Natural Language Processing*.