Improving Statistical Machine Translation using Lexicalized Rule Selection

Zhongjun He[1,2] and **Qun Liu**[1] and **Shouxun Lin**[1]
[1]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, 100190, China
[2]Graduate University of Chinese Academy of Sciences
Beijing, 100049, China
{zjhe,liuqun,sxlin}@ict.ac.cn

Abstract

This paper proposes a novel lexicalized approach for rule selection for syntax-based statistical machine translation (SMT). We build maximum entropy (MaxEnt) models which combine rich context information for selecting translation rules during decoding. We successfully integrate the MaxEnt-based rule selection models into the state-of-the-art syntax-based SMT model. Experiments show that our lexicalized approach for rule selection achieves statistically significant improvements over the state-of-the-art SMT system.

1 Introduction

The syntax-based statistical machine translation (SMT) models (Chiang, 2005; Liu et al., 2006; Galley et al., 2006; Huang et al., 2006) use rules with hierarchical structures as translation knowledge, which can capture long-distance reorderings. Generally, a translation rule consists of a left-hand-side (LHS) [1]and a right-hand-side (RHS). The LHS and RHS can be words, phrases, or even syntactic trees, depending on SMT models. Translation rules can be learned automatically from parallel corpus. Usually, an LHS may correspond to multiple RHS's in multiple rules. Therefore, in statistical machine translation, the rule selection task is to select the correct RHS for an LHS during decoding.

The conventional approach for rule selection is to use precomputed translation probabilities which are estimated from the training corpus, as well as a n-gram language model which is trained on the target language. The limitation of this method is that it ignores context information (especially on the source-side) during decoding. Take the hierarchical model (Chiang, 2005) as an example. Consider the following rules for Chinese-to-English translation [2]:

(1) X → ⟨ 在 $X_{\boxed{1}}$ 的 $X_{\boxed{2}}$, $X_{\boxed{2}}$ in $X_{\boxed{1}}$ ⟩

(2) X → ⟨ 在 $X_{\boxed{1}}$ 的 $X_{\boxed{2}}$, at $X_{\boxed{1}}$'s $X_{\boxed{2}}$ ⟩

(3) X → ⟨ 在 $X_{\boxed{1}}$ 的 $X_{\boxed{2}}$, with $X_{\boxed{2}}$ of $X_{\boxed{1}}$ ⟩

These rules have the same source-side, and all of them can pattern-match all the following source phrases:

(a) 在 [经济 领域]$_1$ 的 [合作]$_2$
in economic field 's cooperation
[cooperation]$_2$ in [the economic field]$_1$

(b) 在 [今天]$_1$ 的 [会议 上]$_2$
at today 's meeting on
at [today]$_1$'s [meeting]$_2$

(c) 在 [人民]$_1$ 的 [支持 下]$_2$
with people 's support under
with [the support]$_2$ of [the people]$_1$

Given a source phrase, how does the decoder know which rule is suitable? In fact, rule (1) and rule (2) have different syntactic structures (the left two trees of Figure 1). Thus rule (1) can be used for translating noun phrase (a), and rule (2) can be applied to prepositional phrase (b). The weakness

[1]In this paper, we use LHS and source-side interchangeably (so are RHS and target-side).

[2]In this paper, we use Chinese and English as the source and target language, respectively.

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 321–328
Manchester, August 2008

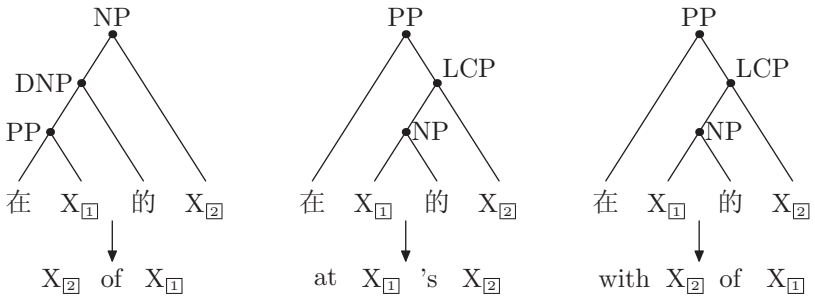

Figure 1: Syntactic structures of the same source-side in different rules.

of Chiang's hierarchical model is that it cannot distinguish different structures on the source-side. The linguistically syntax-based models (Liu et al., 2006; Huang et al., 2006) can distinguish syntactic structures by parsing source sentence. However, as an LHS tree may correspond to different RHS strings in different rules (the right two rules of Figure 1), these models also face the rule selection problem during decoding.

In this paper, we propose a lexicalized approach for rule selection for syntax-based statistical machine translation. We use the maximum entropy approach to combine various context features, e.g., context words of rules, boundary words of phrases, parts-of-speech (POS) information. Therefore, the decoder can use rich context information to perform context-dependent rule selection. We build a maximum entropy based rule selection (MaxEnt RS) model for each *ambiguous hierarchical LHS*, the LHS which contains nonterminals and corresponds to multiple RHS's in multiple rules. We integrate the MaxEnt RS models into the state-of-the-art hierarchical SMT system (Chiang, 2005). Experiments show that the lexicalized rule selection approach improves translation quality of the state-of-the-art SMT system, and the improvements are statistically significant.

2 Previous Work

2.1 The Selection Problem in SMT

Statistical machine translation systems usually face the selection problem because of the one-to-many correspondence between the source and target language. Recent researches showed that rich context information can help SMT systems perform selection and improves translation quality.

The discriminative phrasal reordering models (Xiong et al., 2006; Zens and Ney, 2006) provided a lexicalized method for phrase reordering. In these models, LHS and RHS can be considered as phrases and reordering types, respectively. Therefore the selection task is to select a reordering type for phrases. They use a MaxEnt model to combine context features and distinguished two kinds of reorderings between two adjacent phrases: monotone or swap. However, our method is more generic, we perform lexicalized rule selection for syntax-based SMT models. In these models, the rules with hierarchical structures can handle reorderings of non-adjacent phrases. Furthermore, the rule selection can be considered as a multi-class classification task, while the phrase reordering between two adjacent phrases is a two-class classification task.

Recently, word sense disambiguation (WSD) techniques improved the performance of SMT systems by helping the decoder perform lexical selection. Carpuat and Wu (2007b) integrated a WSD system into a phrase-based SMT system, Pharaoh (Koehn, 2004a). Furthermore, they extended WSD to phrase sense disambiguation (PSD) (Carpuat and Wu, 2007a). Either the WSD or PSD system combines rich context information to solve the ambiguity problem for words or phrases. Their experiments showed stable improvements of translation quality. These are different from our work. On one hand, they focus on solving the lexical ambiguity problem, and use a WSD or PSD system to predict translations for phrases which only consist of words. However, we put emphasis on rule selection, and predict translations for hierarchical LHS's which consist of both words and nonterminals. On the other hand, they incorporated a WSD or PSD system into a phrase-based SMT system with a weak distortion model for phrase reordering. While we incorporate MaxEnt RS models into the state-of-the-art syntax-based SMT system, which captures phrase reordering by using a hierarchical model.

Chan et al. (2007) incorporated a WSD system into the hierarchical SMT system, Hiero (Chiang, 2005), and reported statistically significant improvement. But they only focused on solving ambiguity for terminals of translation rules, and limited the length of terminals up to 2. Different from their work, we consider a translation rule as a whole, which contains both terminals and nonterminals. Moreover, they explored features for the WSD system only on the source-side. While we define context features for the MaxEnt RS models on both the source-side and target-side.

2.2 The Hierarchical Model

The hierarchical model (Chiang, 2005; Chiang, 2007) is built on a weighted synchronous context-free grammar (SCFG) . A SCFG rule has the following form:

$$\text{X} \rightarrow \langle \alpha, \gamma, \sim \rangle \tag{4}$$

where X is a nonterminal, α is an LHS string consists of terminals and nonterminals, γ is the translation of α, $\sim$ defines a one-one correspondence between nonterminals in α and γ. For example,

(5) $\text{X} \rightarrow \langle$ 经济 发展, economic development$\rangle$

(6) $\text{X} \rightarrow \langle \text{X}_{\boxed{1}}$ 的 $\text{X}_{\boxed{2}}$, the $\text{X}_{\boxed{2}}$ of $\text{X}_{\boxed{1}} \rangle$

Rule (5) contains only terminals, which is similar to phrase-to-phrase translation in phrase-based SMT models. Rule (6) contains both terminals and nonterminals, which causes a reordering of phrases. The hierarchical model uses the maximum likelihood method to estimate translation probabilities for a phrase pair $\langle \alpha, \gamma \rangle$, independent of any other context information.

To perform translation, Chiang uses a log-linear model (Och and Ney, 2002) to combine various features. The weight of a derivation D is computed by:

$$w(D) = \prod_i \phi_i(D)^{\lambda_i} \tag{7}$$

where $\phi_i(D)$ is a feature function and λ_i is the feature weight of $\phi_i(D)$. During decoding, the decoder searches the best derivation with the lowest cost by applying SCFG rules. However, the rule selections are independent of context information, except the left neighboring $n-1$ target words for computing n-gram language model.

3 Lexicalized Rule Selection

The rule selection task can be considered as a multi-class classification task. For a source-side, each corresponding target-side is a label. The maximum entropy approach (Berger et al., 1996) is known to be well suited to solve the classification problem. Therefore, we build a maximum entropy based rule selection (MaxEnt RS) model for each ambiguous hierarchical LHS. In this section, we will describe how to build the MaxEnt RS models and how to integrate them into the hierarchical SMT model.

3.1 The MaxEnt RS Model

Following (Chiang, 2005), we use $\langle \alpha, \gamma \rangle$ to represent a SCFG rule extracted from the training corpus, where α and γ are source and target strings, respectively. The nonterminals in α and γ are represented by X_k, where k is an index indicating one-one correspondence between nonterminals in source and target sides. Let us use $f(X_k)$ to represent the source text covered by X_k, and $e(X_k)$ to represent the translation of $f(X_k)$. Let $C(\alpha)$ be the context information of source text matched by α, and $C(\gamma)$ be the context information of target text matched by γ. Under the MaxEnt model, we have:

$$P_{rs}(\gamma|\alpha, f(X_k), e(X_k)) = \frac{\exp[\sum_i \lambda_i h_i(C(\gamma), C(\alpha), f(X_k), e(X_k))]}{\sum_{\gamma'} \exp[\sum_i \lambda_i h_i(C(\gamma'), C(\alpha), f(X_k), e(X_k))]} \tag{8}$$

where h_i is a binary feature function, λ_i is the feature weight of h_i. The MaxEnt RS model combines rich context information of grammar rules, as well as information of the subphrases which will be reduced to nonterminal X during decoding. However, these information is ignored by Chiang's hierarchical model.

We design three kinds of features for a rule $\langle \alpha, \gamma \rangle$:

- Lexical features, which are the words immediately to the left and right of α, and boundary words of subphrase $f(X_k)$ and $e(X_k)$;
- Parts-of-speech (POS) features, which are POS tags of the source words defined in lexical features.
- Length features, which are the length of subphrases $f(X_k)$ and $e(X_k)$.

Side	Type	Name	Description
Source-side	Lexical Features	$W\alpha_{-1}$	The source word immediately to the left of α
		$W\alpha_{+1}$	The source word immediately to the right of α
		$WL_{f(X_k)}$	The first word of $f(X_k)$
		$WR_{f(X_k)}$	The last word of $f(X_k)$
	POS Features	$P\alpha_{-1}$	POS of $W\alpha_{-1}$
		$P\alpha_{+1}$	POS of $W\alpha_{+1}$
		$PL_{f(X_k)}$	POS of $WL_{f(X_k)}$
		$PR_{f(X_k)}$	POS of $WR_{f(X_k)}$
	Length Feature	$LEN_{f(X_k)}$	Length of source subphrase $f(X_k)$
Target-side	Lexical Features	$WL_{e(X_k)}$	The first word of $e(X_k)$
		$WR_{e(X_k)}$	The last word of $e(X_k)$
	Length Feature	$LEN_{e(X_k)}$	Length of target subphrase $e(X_k)$

Table 1: Feature categories of the MaxEnt RS model.

Type	Feature
Lexical Features	$W\alpha_{-1}$=加强 $W\alpha_{+1}$=。
	$WL_{f(X_1)}$=经济 $WR_{f(X_1)}$=领域 $WL_{f(X_2)}$=合作 $WR_{f(X_1)}$=合作
	$WL_{e(X_1)}$=economic $WR_{e(X_1)}$=field $WL_{e(X_2)}$=cooperation $WR_{f(X_1)}$=cooperation
POS Features	$P\alpha_{-1}$=v $W\alpha_{+1}$=wj
	$PL_{f(X_1)}$=n $PR_{f(X_1)}$=n $PL_{f(X_2)}$=vn $PR_{f(X_2)}$=vn
Length Feature	$LEN_{f(X_1)}$=2 $LEN_{f(X_2)}$=1 $LEN_{e(X_1)}$=2 $LEN_{e(X_2)}$=1

Table 2: Features of rule X → ⟨ 在 $X_{[1]}$ 的 $X_{[2]}$, $X_{[2]}$ in the $X_{[1]}$ ⟩.

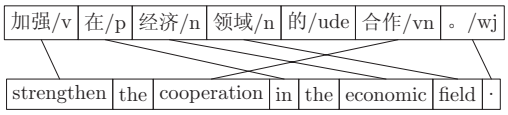

Figure 2: An training example for rule extraction.

Table 1 shows these features in detail.

These features can be easily gathered according to Chinag's rule extraction method (Chiang, 2005). We use an example for illustration. Figure 2 is a word-aligned training example with POS tags on the source side. We can obtain a SCFG rule:

(9) X → ⟨ 在 $X_{[1]}$ 的 $X_{[2]}$, $X_{[2]}$ in the $X_{[1]}$ ⟩

Where the source phrases covered by $X_{[1]}$ and $X_{[2]}$ are "经济 领域" and "合作", respectively. Table 2 shows features of this rule. Note that following (Chiang, 2005), we limit the number of nonterminals of a rule up to 2. Thus a rule may have 20 features at most.

After extracting features from the training corpus, we use the toolkit implemented by Zhang (2004) to train a MaxEnt RS model for each ambiguous hierarchical LHS. We set iteration number to 100 and Gaussian prior to 1.

3.2 Integrating the MaxEnt RS Models into the SMT Model

We integrate the MaxEnt RS models into the SMT model during the translation of each source sentence. Thus the MaxEnt RS models can help the decoder perform context-dependent rule selection during decoding.

In (Chiang, 2005), the log-linear model combines 8 features: the translation probabilities $P(\gamma|\alpha)$ and $P(\alpha|\gamma)$, the lexical weights $P_w(\gamma|\alpha)$ and $P_w(\alpha|\gamma)$, the language model, the word penalty, the phrase penalty, and the glue rule penalty. For integration, we add two new features:

- $P_{rs}(\gamma|\alpha, f(X_k), e(X_k))$. This feature is computed by the MaxEnt RS model, which gives a probability that the model selecting a target-side γ given an ambiguous source-side α, considering context information.
- $P_{rsn} = exp(1)$. This feature is similar to phrase penalty feature. In our experiments,

we find that some source-sides are not ambiguous, and correspond to only one target-side. However, if a source-side α' is not ambiguous, the first feature P_{rs} will be set to 1.0. In fact, these rules are not reliable since they usually occur only once in the training corpus. Therefore, we use this feature to reward the ambiguous source-side. During decoding, if an LHS has multiple translations, this feature is set to $exp(1)$, otherwise it is set to $exp(0)$.

The advantage of our integration is that we need not change the main decoding algorithm of a SMT system. Furthermore, the weights of the new features can be trained together with other features of the translation model.

Chiang (2007) uses the CKY algorithm with a cube pruning method for decoding. This method can significantly reduce the search space by efficiently computing the top-n items rather than all possible items at a node, using the k-best Algorithms of Huang and Chiang (2005) to speed up the computation. In cube pruning, the translation model is treated as the monotonic backbone of the search space, while the language model score is a non-monotonic cost that distorts the search space (see (Huang and Chiang, 2005) for definition of monotonicity). Similarly, in the MaxEnt RS model, source-side features form a monotonic score while target-side features constitute a non-monotonic cost that can be seen as part of the language model.

For translating a source sentence F_I^J, the decoder adopts a bottom-up strategy. All derivations are stored in a *chart* structure. Each cell $c[i,j]$ of the chart contains all partial derivations which correspond to the source phrase f_i^j. For translating a source-side span $[i,j]$, we first select all possible rules from the rule table. Meanwhile, we can obtain features of the MaxEnt RS models which are defined on the source-side since they are fixed before decoding. During decoding, for a source phrase f_i^j, suppose the rule

$$\text{X} \rightarrow \langle f_i^k \text{X}_{\boxed{1}} f_t^j, e_{i'}^{k'} \text{X}_{\boxed{1}} e_{t'}^{j'} \rangle \tag{10}$$

is selected by the decoder, where $i \leq k < t \leq j$ and $k+1 < t$, then we can gather features which are defined on the target-side of the subphrase $\text{X}_{\boxed{1}}$ from the ancestor chart cell $c[k+1, t-1]$ since the span $[k+1, t-1]$ has already been covered. Then the new feature scores P_{rs} and P_{rsn} can be computed. Therefore, the cost of the derivation can be obtained. Finally, the decoding is completed when the whole sentence is covered, and the best derivation of the source sentence F_I^J is the item with the lowest cost in cell $c[I,J]$.

4 Experiments

4.1 Corpus

We carry out experiments on two translation tasks with different sizes and domains of the training corpus.

- IWSLT-05: We use about 40,000 sentence pairs from the BTEC corpus with 354k Chinese words and 378k English words as our training data. The English part is used to train a trigram language model. We use IWSLT-04 test set as the development set and IWSLT-05 test set as the test set.

- NIST-03: We use the FBIS corpus as the training corpus, which contains 239k sentence pairs with 6.9M Chinese words and 8.9M English words. For this task, we train two trigram language models on the English part of the training corpus and the Xinhua portion of the Gigaword corpus, respectively. NIST-02 test set is used as the development set and NIST-03 test set is used as the test set.

4.2 Training

To train the translation model, we first run GIZA++ (Och and Ney, 2000) to obtain word alignment in both translation directions. Then the word alignment is refined by performing "grow-diag-final" method (Koehn et al., 2003). We use the same method suggested in (Chiang, 2005) to extract SCFG grammar rules. Meanwhile, we gather context features for training the MaxEnt RS models. The maximum initial phrase length is set to 10 and the maximum rule length of the source-side is set to 5.

We use SRI Language Modeling Toolkit (Stolcke, 2002) to train language models for both tasks. We use minimum error rate training (Och, 2003) to tune the feature weights for the log-linear model.

The translation quality is evaluated by BLEU metric (Papineni et al., 2002), as calculated by mteval-v11b.pl with case-insensitive matching of n-grams, where $n = 4$.

4.3 Baseline

We reimplement the decoder of Hiero (Chiang, 2007) in C++, which is the state-of-the-art SMT

System	IWSLT-05	NIST-03
Baseline	56.20	28.05
+ MaxEnt RS		
SLex	56.51	28.26
PF	56.95	28.78
SLex+PF	56.99	28.89
SLex+PF+SLen	57.10	28.96
SLex+PF +SLen+TF	57.20	29.02

Table 3: BLEU-4 scores (case-insensitive) on IWSLT-05 task and NIST MT-03 task. SLex = Source-side Lexical Features, PF = POS Features, SLen = Source-side Length Feature, TF = Target-side features.

system. During decoding, we set $b = 100$ to prune grammar rules, $\beta = 10$, $b = 30$ to prune X cells, and $\beta = 10$, $b = 15$ to prune S cells. For cube pruning, we set the threshold $\epsilon = 1.0$. See (Chiang, 2007) for meanings of these pruning parameters.

The baseline system uses precomputed phrase translation probabilities and two trigram language models to perform rule selection, independent of any other context information. The results are shown in the row *Baseline* of Table 3. For IWSLT-05 task, the baseline system achieves a BLEU-4 score of 56.20. For NIST MT-03 task, the BLEU-4 score is 28.05 .

4.4 Baseline + MaxEnt RS

As described in Section 3.2, we add two new features to integrate the MaxEnt RS models into the hierarchical model. To run the decoder, we share the same pruning settings with the baseline system. Table 3 shows the results.

Using all features defined in Section 3.1 to train the MaxEnt RS models, for IWSLT-05 task, the BLEU-4 score is 57.20, which achieves an absolute improvement of 1.0 over the baseline. For NIST-03 task, our system obtains a BLEU-4 score of 29.02, with an absolute improvement of 0.97 over the baseline. Using Zhang's significance tester (Zhang et al., 2004) to perform paired bootstrap sampling (Koehn, 2004b), both improvements on the two tasks are statistically significant at $p < 0.05$.

In order to explore the utility of the context features, we train the MaxEnt RS models on different feature sets. We find that POS features are the most useful features since they can generalize over all training examples. Moreover, length feature also yields improvement. However, these features are never used in the baseline.

	NO. of LHS	NO. of H-LHS	NO. of AH-LHS
NIST MT-03	163,097	148,671	95,424
Baseline	12,069	7,164	5,745
+MaxEnt RS (All features)	12,655	10,306	9,259

Table 4: Number of possible source-sides of SCFG rules for NIST-03 task and number of source-sides of the best translation. H-LHS = Hierarchical LHS, AH-LHS = Ambiguous hierarchical LHS.

5 Analysis

Table 4 shows the number of source-sides of the SCFG rules for NIST-03 task. After extracting grammar rules from the training corpus, there are 163,097 source-sides match the test corpus, 91.15% are hierarchical LHS's (H-LHS, the LHS which contains nonterminals). For the hierarchical LHS's, 64.18% are ambiguous (AH-LHS, the H-LHS which has multiple translations). This indicates that the decoder will face serious rule selection problem during decoding. We also note the number of the source-sides of the best translation for the test corpus. For the baseline system, the number of H-LHS only account for 59.36% of total LHS's. However, by incorporating MaxEnt RS models, that proportion increases to 81.44%, since the number of AH-LHS increases. The reason is that, we use the feature P_{rsn} to reward ambiguous hierarchical LHS's. This has some advantages. On one hand, H-LHS can capture phrase reorderings. On the other hand, AH-LHS is more reliable than non-ambiguous LHS, since most non-ambiguous LHS's occur only once in the training corpus.

In order to know how the MaxEnt RS models improve the performance of the SMT system, we

study the best translation of Baseline and Baseline+MaxEnt RS. We find that the MaxEnt RS models improve translation quality in 2 ways.

5.1 Better Phrase reordering

Since the SCFG rules which contain nonterminals can capture reordering of phrases, better rule selection will produce better phrase reordering. For example, the source sentence "... [联合国 安全 理事会]$_1$ 的 [五 个 常任 理事国]$_2$ ... " is translated as follows:

- Reference: ... *the five permanent members of the UN Security Council* ...
- Baseline: ... *the [United Nations Security Council]$_1$ [five permanent members]$_2$* ...
- +MaxEnt RS: ... *[the five permanent members]$_2$ of [the UN Security Council]$_1$* ...

The source sentence is translated incorrectly by the baseline system, which selects the rule

(11) X → ⟨ X$_{\boxed{1}}$ 的 X$_{\boxed{2}}$, the X$_{\boxed{1}}$X$_{\boxed{2}}$ ⟩

and produces a monotone translation. In contrast, by considering information of the subphrases X$_{\boxed{1}}$ and X$_{\boxed{2}}$, the MaxEnt RS model chooses the rule

(12) X → ⟨ X$_{\boxed{1}}$ 的 X$_{\boxed{2}}$, X$_{\boxed{2}}$ of X$_{\boxed{1}}$ ⟩

and obtains a correct translation by swapping X$_{\boxed{1}}$ and X$_{\boxed{2}}$ on the target-side.

5.2 Better Lexical Translation

The MaxEnt RS models can also help the decoder perform better lexical translation than the baseline. This is because the SCFG rules contain terminals. When the decoder selects a rule for a source-side, it also determines the translations of the source terminals. For example, the translations of the source sentence "恐怕 这 趟 航班 已经 订 满 了 。" are as follows:

- Reference *I'm afraid this flight is full.*
- Baseline: *I'm afraid already booked for this flight.*
- +MaxEnt RS: *I'm afraid this flight is full.*

Here, the baseline translates the Chinese phrase "订 满" into "booked" by using the rule:

(13) X → ⟨ X$_{\boxed{1}}$ 订 满, X$_{\boxed{1}}$ booked⟩

The meaning is not fully expressed since the Chinese word "满" is not translated. However, the MaxEnt RS model obtains a correct translation by using the rule:

(14) X → ⟨ X$_{\boxed{1}}$ 订 满, X$_{\boxed{1}}$ full ⟩

However, we also find that some results produced by the MaxEnt RS models seem to decrease the BLEU score. An interesting example is the translation of the source sentence "这 条 街 叫 什 么 ？ ":

- Reference1: *What is the name of this street?*
- Reference2: *What is this street called?*
- Baseline: *What is the name of this street?*
- +MaxEnt RS: *What's this street called?*

In fact, both translations are correct. But the translation of the baseline fully matches *Reference1*. Although the translation produced by the MaxEnt RS model is almost the same as *Reference2*, as the BLEU metric is based on n-gram matching, the translation "What's" cannot match "What is" in *Reference2*. Therefore, the MaxEnt RS model achieves a lower BLEU score.

6 Conclusion

In this paper, we propose a generic lexicalized approach for rule selection. We build maximum entropy based rule selection models for each ambiguous hierarchical source-side of translation rules. The MaxEnt RS models combine rich context information, which can help the decoder perform context-dependent rule selection during decoding. We integrate the MaxEnt RS models into the hierarchical SMT model by adding two new features. Experiments show that the lexicalized approach for rule selection achieves statistically significant improvements over the state-of-the-art syntax-based SMT system.

Furthermore, our approach not only can be used for the formally syntax-based SMT systems, but also can be applied to the linguistically syntax-based SMT systems. For future work, we will explore more sophisticated features for the MaxEnt RS models and integrate the models into the linguistically syntax-based SMT systems.

Acknowledgements

We would like to show our special thanks to Hwee Tou Ng, Liang Huang, Yajuan Lv and Yang Liu for their valuable suggestions. We also appreciate the anonymous reviewers for their detailed comments and recommendations. This work was supported by the National Natural Science Foundation of China (NO. 60573188 and 60736014), and the High Technology Research and Development Program of China (NO. 2006AA010108).

References

Berger, A. L., S. A. Della Pietra, and V. J. Della. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, page 22(1):39 – 72.

Carpuat, Marine and Dekai Wu. 2007a. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *11th Conference on Theoretical and Methodological Issues in Machine Translation*, pages 43–52.

Carpuat, Marine and Dekai Wu. 2007b. Improving statistical machine translation using word sense disambiguation. In *Proceedings of EMNLP-CoNLL 2007*, pages 61–72.

Chan, Yee Seng, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 33–40.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270.

Chiang, David. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, pages 33(2):201–228.

Galley, Michel, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of COLING-ACL 2006*, pages 961–968.

Huang, Liang and David Chiang. 2005. Better k-best parsing. In *Proceedings of the 9th International Workshop on Parsing Technologies*.

Huang, Liang, Kevin Knight, and Aravind Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*.

Koehn, Philipp, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133.

Koehn, Philipp. 2004a. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas*, pages 115–124.

Koehn, Philipp. 2004b. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

Liu, Yang, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pages 609–616.

Och, Franz Josef and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Stolcke, Andreas. 2002. Srilm – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken language Processing*, volume 2, pages 901–904.

Xiong, Deyi, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pages 521–528.

Zens, Richard and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 55–63.

Zhang, Ying, Stephan Vogel, and Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2051–2054.

Zhang, Le. 2004. Maximum entropy modeling toolkit for python and c++. available at http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.