A Framework for Discriminative Rule Selection in Hierarchical Moses

Fabienne Braune[1], Alexander Fraser[1], Hal Daumé III[2], and Aleš Tamchyna[3]

[1]Center for Information and Language Processing, University of Munich, Germany
[2]Computer Science and UMIACS, University of Maryland
[3]Charles University in Prague, Prague Czech Republic

Abstract

Training discriminative rule selection models is usually expensive because of the very large size of the hierarchical grammar. Previous approaches reduced the training costs either by (i) using models that are local to the source side of the rules or (ii) by heavily pruning out negative samples. Moreover, all previous evaluations were performed on small scale translation tasks, containing at most 250,000 sentence pairs. We propose two contributions to discriminative rule selection. First, we test previous approaches on two French-English translation tasks in domains for which only limited resources are available and show that they fail to improve translation quality. To improve on such tasks, we propose a rule selection model that is (i) global with rich label-dependent features (ii) trained with all available negative samples. Our global model yields significant improvements, up to 1 BLEU point, over previously proposed rule selection models. Second, we successfully scale rule selection models to large translation tasks but have so far failed to produce significant improvements in BLEU on these tasks.

1 Introduction

Hierarchical phrase-based machine translation (Chiang, 2005) performs non-local reordering in a formally syntax-based way. It allows flexible rule extraction and application by using a grammar without linguistic annotation. As a consequence, many hierarchical rules can be used to translate a given input segment even though only a subset of these yield a correct translation. For instance, rules r_1 to r_3 can be applied to translate the French sentence F_1 below although only r_1 yields the correct translation E.

(r_1) X $\rightarrow \langle$ X_1 pratique X_2, practical X_1 X_2 $\rangle$
(r_2) X $\rightarrow \langle$ X_1 pratique X_2, X_1 X_2 practice $\rangle$
(r_3) X $\rightarrow \langle$ X_1 pratique X_2, X_2 X_1 process $\rangle$

F_1 Une étude de l' (intérêt)$_{X_1}$ **pratique** (de notre approche)$_{X_2}$.
A study on the (interest)$_{X_1}$ ***practical*** *(of our approach)$_{X_2}$.*

E A study on the **practical** (interest)$_{X_1}$ (of our approach)$_{X_2}$.

The rule scoring heuristics defined by (Chiang, 2005) do not handle rule selection in a satisfactory way and many authors have come up with solutions. Models that use the syntactic structure of the source and target sentence have been proposed by (Marton and Resnik, 2008; Marton et al., 2012; Chiang et al., 2009; Chiang, 2010; Liu et al., 2011). These approaches exclusively take into account syntactic structure and do not model rule selection (see Section 6 for a detailed discussion). Following the work on phrase-sense disambiguation by (Carpuat and Wu, 2007), other authors improve rule selection by defining features on the structure of hierarchical rules and combining these with information about the source sentence (Chan et al., 2007; He et al., 2008; He et al., 2010; Cui et al., 2010). In these approaches, rule selection is the task of selecting the target side of a rule given its source side as well as contextual information about the source sentence. This task is modeled as a multiclass classification problem.

Because of the very large size of hierarchical grammars, the training procedure for discriminative rule selection models is typically very expensive: multiclass classification is performed over

Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers, pages 92–101,
Berlin, Germany, August 11-12, 2016.

millions of classes (one for each possible target side of a hierarchical rule). To overcome this problem, previous approaches reduced the training costs by either (i) using models that are local to the source side of hierarchical rules or (ii) heavily pruning out negative samples from the training data. (Chan et al., 2007; He et al., 2008; He et al., 2010) train one (local) classifier for each source side or pattern of hierarchical rules instead of defining a (global) model over all rules. Cui et al. (2010) train global models but in addition to rule table pruning, they heavily prune out negative instances. Finally, in all previous approaches, a small amount of fixed features is used for training and prediction.

While previous approaches have been shown to work on a small[1] English-Chinese news translation task, we show (in Section 4) that on French-English tasks on domains for which only a limited amount of training data is available (which we call low resource tasks), they fail to improve over a hierarchical baseline. This failure is caused by the fact that the models proposed so far do not take advantage of all information available in the training data. Local models prevent feature sharing between rules with different source sides or patterns (see Section 2.3) while aggressive pruning removes important information from the training data (see Section 3.2). On low resource translation tasks, this loss hurts translation quality. Moreover, the small set of features used in previous work does not provide a representation of the training data that is as powerful as it could be for classification (see Section 2.2).

We improve on previous work in two ways. First, we define a global rule selection model with a rich set of feature combinations. Our global model enables feature sharing while the large amount of features we use offers a complete representation of the available training data. We train our model with all acquired training examples. The exhaustive training of a feature rich global model allows us to take full advantage of the training data. We show on two low-resource French-English translation tasks that local and pruned models often fail to improve over a hierarchical baseline while our global model with exhaustive training yields significant improvements on scientific and medical texts (see Section 4). In a second contribution, we successfully scale rule selection models to large scale translation tasks but fail to produce significant improvements in BLEU over a hierarchical baseline on these tasks.

Because our approach needs scaling to a large amount of training examples, we need a classifier that is fast and supports online streaming. We use the high-speed classifier Vowpal Wabbit[2] (VW) which we fully integrate in the syntax component (Hoang et al., 2009) of the Moses machine translation toolkit (Koehn et al., 2007). To allow researchers to replicate our results and improve on our work, we make our implementation publicly available as part of Moses.

2 Global Rule Selection Model

The goal of rule selection is to choose the correct target side of a hierarchical rule, given a source side as well as other sources of information such as the shape of the rule or its context of application in the source sentence. The latter includes lexical features (e.g. the words surrounding the source span of an applied rule) or syntactic features (e.g. the position of an applied rule in the source parse tree). The rule selection task can be modeled as a multi-class classification problem where each target-side corresponding to a source side gets a label.

Contrary to (Chan et al., 2007; He et al., 2008; He et al., 2010), we solve the classification problem by building a single global discriminative model instead of using one maximum entropy classifier for each source side or pattern. We solve the rule selection problem through multi-class classification while (Cui et al., 2010) approximate the problem by using a binary classifier.

2.1 Model Definition

We denote SCFG rules by $X \rightarrow \langle \alpha, \gamma \rangle$, where α is a source and γ a target language string (Chiang, 2005). By $C(f, \alpha)$ we denote information of the source sentence f and the source side α. $R(\alpha, \gamma)$ represents features on hierarchical rules. Our discriminative model estimates $P(\gamma \mid \alpha, C(f, \alpha), R(\alpha, \gamma))$ and is normalized over the set G' of candidate target sides γ' for a given α. The function $GTO : \alpha \rightarrow G'$ generates, given the source side, the set G' of all corresponding target sides γ'. The estimated distribution can be written

[1] In (He et al., 2008; Cui et al., 2010), the size of the training data is about 240k parallel sentences.

[2] `http://hunch.net/~vw/`. Implemented by John Langford and many others.

as:

$$P(\gamma \mid \alpha, C(f,\alpha), R(\alpha,\gamma)) = \frac{\exp(\sum_i \lambda_i h_i(\alpha, C(f,\alpha), R(\alpha,\gamma)))}{\sum_{\gamma' \in GTO(\alpha)} \exp(\sum_i \lambda_i h_i(\alpha, C(f,\alpha), R(\alpha,\gamma')))}$$

In the same fashion as for local models, our global model predicts the target side of a rule given its source side and contextual features, meaning that it still disambiguates between rules with the *same* source side using rich context information. However, because the global model trains a *single* classifier over all rules, it captures information that is shared among rules with different source sides (see Section 2.3 for more details).

2.2 Feature Templates

We now present the feature templates $R(\alpha,\gamma)$ and $C(f,\alpha)$ in the equation presented in Section 2.1. While in isolation the features composing the templates are similar to the features used in previous work (He et al., 2008; He et al., 2010; Cui et al., 2010), we create powerful representations by dividing our feature set into fixed and label-dependent features and taking the cross product of these.

We begin by presenting the features in our templates. To this aim suppose that rule r_4 has been extracted from sentence F_2. The 1-best parse tree of F_2 is given in Figure 1.

(r_4) X $\rightarrow \langle$ pratique X_1 X_2, X_2 X_1 process $\rangle$

F_2 Une étude de la **pratique** (de l'ingénérie)$_{X_1}$ (informatique)$_{X_2}$
*A study on the **process** (of software)$_{X_1}$ (development)$_{X_2}$.*

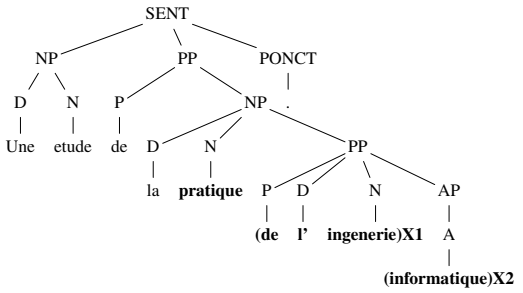

Figure 1: Parse tree of Sentence F_2

The rule internal features $R(\alpha,\gamma)$ are given in Figure 2. The source context features $C(f,\alpha)$ are divided into (i) lexical and (ii) syntactic features. Lexical features are given in Figure 3 where the term "factored form" denotes the surface form, POS tag and lemma of a word. Syntactic features are in Figure 4.

In order to create powerful representations, we combine the features above into more complex templates. To this aim, we distribute our features into two categories:

1. A set of fixed features S on the source sentence context and source side of the rule.
2. A set of features T which varies with the target side of the rule, which we call *label-dependent*.

The set S includes the lexical and syntactic features in Figures 3 and 4 as well as shape features on the source side α (2 first rows of Figure 2). The set T contains all shape features involving the target side of the rules (5 last rows of Figure 2). Our feature space consists of all source and target features S and T as well as the cross product $S \times T$.

The features resulting from the cross product $S \times T$ capture many aspects of rule selection that are lost when the features are considered in isolation. For instance, the cross product of (i) the lexical features (Figure 3) and source word shape features (Figure 2, row 2) with (ii) the target word shape features (Figure 2, row 4) create typical templates of a discriminative word lexicon. In the same fashion, the cross product of (i) the syntactic features (Figure 4) with (ii) the target alignment shape feature (Figure 2, row 6) creates the templates of a reordering model using syntactic features.

2.3 Feature Sharing

An advantage of global models over local ones is that they allow feature sharing between rules with different source sides. Through sharing, features that do not depend on the source side of rules but are nevertheless often seen across all rules can be captured. As an illustration, assume that rules r_5 and r_6 have been extracted from sentence F_3 below. The 1-best parse of F_3 is given in Figure 5.

(r_5) X $\rightarrow \langle$ modèles X_1 de bas X_2, X_1 X_2 models $\rangle$
(r_6) X $\rightarrow \langle$ modèles X_1 de X_2, X_1 X_2 models $\rangle$

F_3 Un article sur les **modèles** (statistiques)$_{X_1}$ de (bas niveau)$_{X_2}$.

Feature Template	Example
Source side α	*pratique_X1_X2* (one feature)
Words in α	*pratique X1 X2* (three features)
Target side γ	*X2_X1_process*
Words in γ	*X2 X1 process*
Aligned terminals in α and γ	*pratique↔process*
Aligned non-terminals in α and γ	*X1↔X2 X2↔X1* (two features)
Best baseline translation probability	*Most_Frequent*

Figure 2: Rule shape features

Feature Template	Example
first factored form left of α	*la, D, la*
second factored form left of α	*de, P, de*
first factored form right of α	*., PONCT, .*
second factored form right of α	*None, None, None*

Figure 3: Lexical features

Feature Template	Example
Does α match a constituent	*no_match*
Type of matched constituent	*None*
Parent of matched constituent	*None*
Lowest parent of unmatched constituent	*NP*
Span width covered by α	*5*

Figure 4: Syntactic features

A paper on the ***models*** *(statistical)$_{X_1}$ of (low-level)$_{X_2}$*

Although r_4, r_5 and r_6 have completely different source sides, they share many contextual features such as:

(i) The POS tags of the first and second words to the left of the segment where the rules are applied (which are P and D)
(ii) The syntactic structure of this segment (which is that (i) it is not a complete constituent and (ii) it has a NP as its lowest parent)
(iii) The rule span width (which is 5)

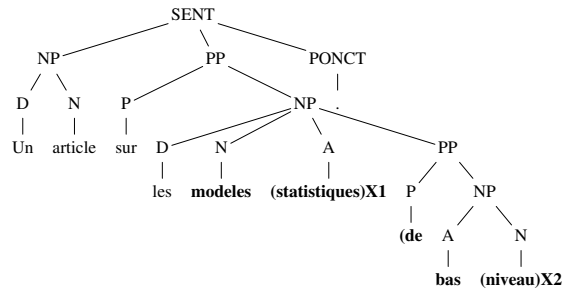

Figure 5: Parse tree of Sentence F_3

A global model would assign high weights to features (i) to (iii) while local models fail to capture this generalization.

3 Exhaustive Model Training

Training examples for our classifier are generated each time a hierarchical rule can be extracted from the parallel corpus (see Section 3.1). This procedure leads to a very large number of training examples. In contrast to (Cui et al., 2010), we do not prune out negative samples and use all available data to train our model.

3.1 Training procedure

We create training examples using the rule extraction procedure in (Chiang, 2005). We first extract a rule-table in the standard way. Then, each time a rule $a_1 : X \rightarrow \langle \alpha, \gamma \rangle$ can be extracted from the parallel corpus, we create a new training example. γ is the correct class and receives a cost of 0. We create incorrect classes using the rules $a_2, \ldots, a_n$ in the rule-table that have the same source side as a_1 but different target sides. As an example, suppose that rule r_1 introduced in Section 1 has been extracted from sentence F_1. The target side "practical X_1 X_2" is a correct class and gets a cost

of 0. The target side of all other rules having the same source side, such as r_2 and r_3, are incorrect classes.

This process leads to a very large number of training examples, and for each of these we generally have multiple incorrect classes. The total number of training examples for our French-English data sets are displayed in Table 1. We do

Data	Science	Medical	News
Sentences	139,215	111,165	1,572,099
Examples	47,952,867	25,435,958	583,165,140
cost 0	50,718,190	26,458,411	597,575,905
cost 1	493,271,397	170,064,556	8,805,099,861
avg 1	10.28	6.68	15.09

Table 1: Number of training examples (Examp.) The last line shows the average amount of negative samples (avg 1) for each training example.

not prune out negative instances and use all acquired examples for model training. To scale to this amount of training samples, we use the high-speed classifier Vowpal Wabbit (VW). For model training, we use the cost-sensitive one-against-all-reduction (Beygelzimer et al., 2005) of VW. Specifically, the training algorithm which we use is the label dependent version of Cost Sensitive One Against All which uses classification.[3] Two features of VW which are useful for our work are feature hashing and quadratic feature expansion. The quadratic expansion allows us to take the cross-product of the simple source and target features without having to actually write this expansion to disk, which would be prohibitive. Feature hashing (Weinberger et al., 2009) is also important for scaling the classifier to the enormous number of features created by the cross-product expansion.

We avoid overfitting to training data by employing early stopping once classifier accuracy decreases on a held-out dataset.[4] Our model is integrated in the hierarchical framework as an additional feature of the log-linear model.

3.2 Training without Pruning of Negative Examples

By not pruning negative samples, we keep important information for model training. As an illustration, consider the example presented above (Section 3.1) where rule r_1 is a positive instance and r_2 and r_3 are negative samples. The negative instances indicate that in the context of sentence F_1, the internal features of r_2 and r_3 are not correct. For instance, a piece of information that could be paraphrased into I is lost.

I In the syntactic and lexical context of F_1 the terminal *pratique* should neither be translated into *practice* nor into *process*

Consider sentence F_4, which has a similar context to F_1 in terms of the lexical and syntactic features described in Section 2.2. To illustrate the syntactic features common to F_1 and F_4, we give the 1-best parse trees of these sentences in Figures 6 and 7.

F_4 Les avantages de l' (aspect)$_{X_1}$ **pratique** (de la robotique)$_{X_2}$.
The advantages of the (aspect)$_{X_1}$ ***practical*** *(of robotics)$_{X_2}$.*

In pruning-based approaches, if r_2 and r_3 appear infrequently in the training data, they are pruned out and information I is lost. If at decoding time candidate rules that share features with r_2 and r_3 are bad candidates to translate F_1 and F_4 then their application is not blocked by the discriminative model basing on I. For instance, if rules r_7 and r_8 have high scores in the hierarchical model but are bad candidates in the context of sentences F_1 and F_4 then a pruned model fails to block their application. In other words, the discriminative model does not learn that rules containing the lexical items *practice* and *process* on the target language side are bad candidates to translate F_1 and F_4. As a consequence, the application of r_7 and r_8 to F_4 generates the erroneous translations E_1^* and E_2^* below.

(r_7) X $\rightarrow \langle$ X_1 pratique X_2, X_2 X_1 practice $\rangle$
(r_8) X $\rightarrow \langle$ X_1 pratique X_2, X_1 X_2 process $\rangle$

E_1^* The advantages of the of robotics aspects practice
E_2^* The advantages of the aspects of robotics process

4 Experiments on small domains

In a first set of experiments, we evaluate our approach on two low resource French-English trans-

[3] The command line parameter to VW is "csoaa_ldf mc".

[4] We use the development set which is also used for tuning with MIRA, as we will discuss later in the paper.

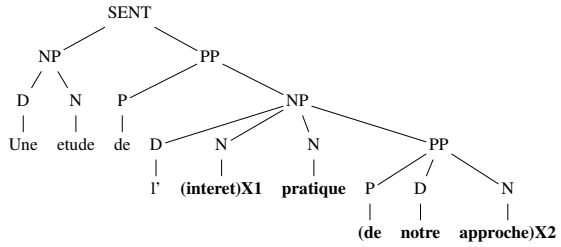

Figure 6: Parse tree of Sentence F_1

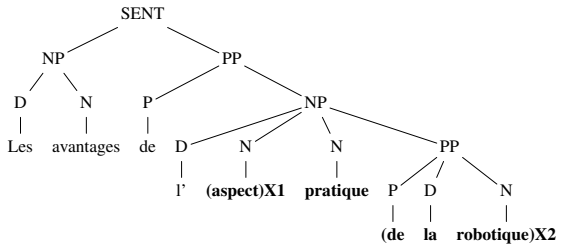

Figure 7: Parse tree of Sentence F_4

lation tasks: (i) a set of scientific articles and (ii) a set of biomedical texts. As these data sets cover small domains, they allow us to investigate the usefulness of our approach in this context. The goal of our experiments is to verify three hypotheses:

- h_1 *Our approach beats a hierarchical baseline.*
- h_2 *Our global model outperforms its local variants.*
- h_3 *Our exhaustive training procedure beats systems trained with pruned data.*

4.1 Experimental Setup

Our scientific data consists of the scientific abstracts provided by Carpuat et al. (2013). The training data contains 139,215 French and English parallel sentences. The development and test sets both consist of 3916 parallel sentences. For the medical domain, we use the biomedical data from EMEA (Tiedemann, 2009). As training data, we used 472,231 sentence pairs from EMEA. We removed duplicate sentences and constructed development and test data by randomly selecting 4000 sentence-pairs. After removal of duplicate sentences, development and test data, we obtain 111,165 parallel sentences for training. For all data sets, we trained a 5-gram language model using the SRI Language Modeling Toolkit (Stolcke, 2002). The training data for the language model is the English side of the training corpus for each task.

We train the model in the standard way, using GIZA++. After training, we reduce the number of translation rules using significance testing (Johnson et al., 2007). For feature extraction, we parse the French part of our training data using the Berkeley parser (Petrov et al., 2006) and lemmatize and POS tag it using Morfette (Chrupała et al., 2008). We train the rule-selection model using VW. All systems are tuned using batch MIRA (Cherry and Foster, 2012). We measure the overall translation quality using 4-gram BLEU (Papineni et al., 2002), which is computed on tokenized and lowercased data for all systems. Statistical significance is computed with the pairwise bootstrap resampling technique of (Koehn, 2004).

4.2 Compared Systems

We investigate systems including a discriminative model in the three setups, given in Figure 4.2. For each setup, we train a global model using a single classifier. For instance, for the setup (LexGlob) we train a classifier with the lexical and rule shape features presented in Section 2.2 together with their cross product.

Description	Name
Rule shape and lexical features	LexGLob
Rule shape and syntactic features	SyntGlob
Rule shape, lexical and syntactic features	LexSyntGlob

Figure 8: Setups of evaluated discriminative models.

In order to verify our first hypothesis (h_1), we show that our approach yields significant improvements over the hierarchical model in (Chiang, 2005). The results of this experiment are given in Table 2.

To verify our second hypothesis (h_2), we show that global rule selection models significantly improve over their local variants. For this second evaluation, we train local models with the feature templates in Figure 4.2. Local models with rule shape and lexical features are used in (He et al., 2008). We further test the performance of local rule selection models by also including syntactic features and a combination of those with the lexical features. We report the results in Table 3 where the local systems are denoted by *LexLoc*, *SyntLoc* and *LexSyntLoc*.

For our third hypothesis (h_3), we show that pruning hurts translation quality. To this aim, we take our best performing global model, which

uses syntactic and rule shape features and perform heavy pruning of negative examples in the data used for classifier training. To exactly reproduce the context-based target model in (Cui et al., 2010), we pruned as many negative examples as necessary to obtain approximately the same amount of positive and negative examples they report. We removed negative instances created from rules with target side frequency < 5000. In the next section, we denote this system by *SyntPrun* and compare it to the hierarchical baseline as well as to our global model in Table 4.

4.3 Results

The outcome of our experiments confirm hypotheses h_1 and h_3 on all data sets and h_2 on medical data only.

The results of our first evaluation (Table 2) show that on all data sets our global rule selection model outperforms the hierarchical baseline (h_1).

The results of our second evaluation (i.e. local vs. global models in Table 3) show that h_2 holds on the medical domain only. On scientific data, global rule selection models in all setups perform slightly better than their local versions but the difference is not statistically significant. Note that all rule selection models except *LexLoc* outperform the hierarchical baseline. The best performing system is a global model with syntactic features (*SyntGlob*). On medical texts, global models outperform their local variants for all feature templates. In each setup, the improvement of local models over the global ones is statistically significant. *SyntGlob* achieves the best performance and yields significant improvements over the baseline. The good performance of *SyntGlob* on scientific and especially medical data can be explained by the fact that syntactic features are less sparse than lexical features and hence generalize better. This is especially important within a global model that allows feature sharing between source sides of rules. Even a combination of lexical and syntactic features underperforms syntactic features on their own because of the sparse lexical features.

The results of our third evaluation are displayed in Table 4. These show that on all data sets our global model without pruning outperforms the same model with pruned training data (h_3). These results also show that the pruned model fails to outperform the hierarchical baseline. Note that this result is consistent with the results reported in (Cui et al., 2010): their Context-based target model yields very low improvements when used in isolation.

System	Science	Medical
Hierarchical	31.22	48.67
LexGlob	**31.69**	48.94
LexSyntGlob	**31.89**	48.97
SyntGlob	**32.27**	**49.66**

Table 2: Evaluation of global models against hierarchical baseline. The results in bold are statistically significant improvements over the Baseline (at confidence $p < 0.05$).

System	Science	Medical
Hierarchical	31.22	48.67
LexLoc	31.50	48.43
LexSyntLoc	**31.74**	48.51
SyntLoc	**31.85**	48.76
LexGlob	**31.69**	48.94*
LexSyntGlob	**31.89**	48.97*
SyntGlob	**32.27**	**49.66***

Table 3: Evaluation of global models against local. We use * to mark global systems that yield statistically significant (at confidence $p < 0.05$) improvements over their local variants. The results in bold are statistically significant improvements over the hierarchical baseline.

5 Large scale Experiments

In a second set of experiments, we evaluate the usefulness of our approach on two large scale translation tasks: (i) a French-to-English news translation task trained on 1,500,000 parallel sentences and (ii) an English-to-Romanian news translation task trained on 600,000 parallel sentences. The training data for the first task consists of the French-English part of the Europarl-v4 corpus. Development and test sets are from the French-to-English news translation task of WMT 2009 (Callison-Burch et al., 2009). For the second task, we use the English-Romanian part of the Europarl-v8 corpus. Development and test sets are from the English-to-Romanian news translation task of WMT 2016. The setup of these experiments is the same as described in Section 4.1 except for the language model of the English-to-Romanian task, which was trained using lmplz

System	Science	Medical
Hierarchical	31.22	48.67
SyntGlob	**32.27**	**49.66**
SyntPrun	31.00	48.61

Table 4: Evaluation of global model against pruned. The results in bold are statistically significant improvements over the Baseline (at confidence $p < 0.05$).

System	Fr-En News	En-Ro News
Hierarchical	20.96	24.16
LexGlob	21.01	24.23
LexSyntGlob	21.04	24.19
SyntGlob	21.14	24.52

Table 5: Evaluation of large scale tasks. No significant difference in performance between the evaluated models.

(Heafield et al., 2013) on the Romanian part of the Common Crawl corpus.

Our goal is to verify if on large scale translation tasks our global rule selection model outperforms a hierarchical baseline (hypothesis h_1 above). The results, given in Table 5, show that on large scale tasks, rule selection models with syntactic features yield small improvements over the hierarchical baseline. However, none of these is statistically significant. Hence hypothesis h_1 does not hold on large domains.

6 Related Work

(Marton and Resnik, 2008; Marton et al., 2012) improve hierarchical machine translation by augmenting the translation model with fine-grained syntactic features of the source sentence. The used features reward rules that match syntactic constituents and punish non-matching rules. (Chiang et al., 2009) integrate these features into a translation model containing a large number of other features such as discount or insertion features. (Chiang, 2010) extends the approach in (Marton and Resnik, 2008) by also including syntactic information of the target sentence that is built during decoding while (Liu et al., 2011) define a discriminative model over source side constituent labels instead of rewarding matching constituents. The training data for their model is based on source sentence derivations.[5] In contrast to this work, we define a rule selection model, i.e. a discriminative model on the target side of hierarchical rules. The training data for our model is based on the hierarchical rule extraction procedure: we acquire training instances by labeling candidate rules extracted from the same sentence pairs.

Similar to our work, (He et al., 2008) define a discriminative rule selection model including lexical features, similar to the ones we presented in Section 2.2. Their work bases on (Chan et al., 2007) which integrate a word sense disambiguation system into a hierarchical system. As opposed to (He et al., 2008), this work focuses on hierarchical rules containing only terminal symbols and having length 2. These approaches train rule selection models that are local to the source side of hierarchical rules. (He et al., 2010) generalize this work by defining a model that is local to source patterns instead of the source side of each rule. We extend these approaches by defining a global model that generalizes to all rules instead of rules with the same source side or source pattern. We also extend the feature set by defining models on syntactic features.

(Cui et al., 2010) propose a joint rule selection model over the source and target side of hierarchical rules. Our work is similar to their Context Based Target Model (CBTM) but it integrates much more information by not reducing the rule selection problem to a binary classification problem and by not pruning the set of negative examples. We show empirically that the exhaustive training of our model significantly improves over their CBTM.

Finally, several authors train local rule selection models for different types of syntax- and semantics- based systems. (Liu et al., 2008) train a local discriminative rule selection model for tree-to-string machine translation. (Zhai et al., 2013) propose a discriminative model to disambiguate predicate argument structures (PAS). In contrast, our rule selection model uses syntactic features on hierarchical rules and is a global model.

All[6] of the mentioned models are trained using the maximum entropy approach (Berger et al., 1996) which seems not to scale well as reported in

[5] The training instances are obtained by performing bilingual parsing on the training data and extracting the obtained rules from the derivation forest.

[6] All of the models except (Chan et al., 2007) which uses an SVM, which is also not efficient.

(Cui et al., 2010). By using a high-speed streaming classifier we are able to train a global model doing true multi-class classification without pruning of training examples.

7 Conclusion and Future Work

We have presented two contributions to previous work on rule selection. First, we improved translation quality on low resource translation tasks by defining a global discriminative rule selection model trained on all available training examples. In a second contribution, we successfully scaled our global rule selection model to large scale translation tasks and presented the first evaluation of discriminative rule selection on such tasks. However, we failed so far to produce significant improvements in BLEU over a hierarchical baseline on large scale French-to-English and English-to-Romanian translation tasks. To allow researchers to replicate our results and improve on our work, we make our implementation publicly available as part of Moses.

Acknowledgements

We thank all members of the DAMT team of the 2012 JHU Summer Workshop. This work was partially supported by Deutsche Forschungsgemeinschaft grant Models of Morphosyntax for Statistical Machine Translation (Phase 2). This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644402 (HimL) and from the European Research Council (ERC) under grant agreement No. 640550.

References

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

Alina Beygelzimer, John Langford, and Bianca Zadrozny. 2005. Weighted one-against-all. In *AAAI*, pages 720–725.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. 4th Workshop on Statistical Machine Translation*, pages 1–28.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *In The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007*, pages 61–72.

Marine Carpuat, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013. SenseSpotting: Never let your parallel data tie you to an old domain. In *Proc. ACL*.

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

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *NAACL 2012*.

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proc. NAACL*.

David Chiang. 2005. Hierarchical phrase-based translation. In *Proceedings of the 43th Annual Meeting of the Association of Computational Linguistics*, page 263270. Association for Computational Linguistics.

David Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics*, pages 1443–1452. The Association for Computer Linguistics.

Grzegorz Chrupała, Georgiana Dinu, and Josef Van Genabith. 2008. Learning morphology with morfette. In *LREC 2008*.

Lei Cui, Dongdong Zhang, Mu Li, Ming Zhou, and Tiejun Zhao. 2010. A joint rule selection model for hierarchical phrase-based translation. In *Proceedings of the ACL 2010 Conference*, pages 6–11. Association for Computational Linguistics.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *COLING*, pages 321–328.

Zhongjun He, Yao Meng, and Hao Yu. 2010. Maximum entropy based phrase reordering for hierarchical phrase-based translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proc. ACL*.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT*, pages 152–159.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL: Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. ACL.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*, pages 388–395.

Qun Liu, Zhongjun He, Yang Liu, and Shouxun Lin. 2008. Maximum entropy based rule selection model for syntax-based statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 89–97. Association for Computational Linguistics.

Lemao Liu, Tiejun Zhao, Chao Wang, and Hailong Cao. 2011. A unified and discriminative soft syntactic constraint model for hierarchical phrase-based translation. In *Proceedings of the 13th Machine Translation Summit*, pages 253–261.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, pages 1003–1011.

Yuval Marton, David Chiang, and Philip Resnik. 2012. Soft syntactic constraints for arabic—english hierarchical phrase-based translation. *Machine Translation*, 26(1-2):137–157, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th ACL*, pages 311–318.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics.

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

Jörg Tiedemann. 2009. News from opus : A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing V*, volume V, pages 237–248. John Benjamins.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. 2009. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*.

Feifei Zhai, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2013. Handling ambiguities of bilingual predicate-argument structures for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1127–1136, Sofia, Bulgaria, August. Association for Computational Linguistics.