Rule Selection with Soft Syntactic Features for String-to-Tree Statistical Machine Translation

Fabienne Braune and Nina Seemann and Alexander Fraser
CIS, Ludwig-Maximilians-Universität München
Oettingenstraße 67, 80538 München, Germany
[braunefe|seemanna]@ims.uni-stuttgart.de
fraser@cis.uni-muenchen.de

Abstract

In syntax-based machine translation, rule selection is the task of choosing the correct target side of a translation rule among rules with the same source side. We define a discriminative rule selection model for systems that have syntactic annotation on the target language side (string-to-tree). This is a new and clean way to integrate soft source syntactic constraints into string-to-tree systems as features of the rule selection model. We release our implementation as part of Moses.

1 Introduction

Syntax-based machine translation is well known for its ability to handle non-local reordering. Syntax-based models either use linguistic annotation on the source language side (Huang, 2006; Liu et al., 2006), target language side (Galley et al., 2004; Galley et al., 2006) or are syntactic in a structural sense only (Chiang, 2005). Recent shared tasks have shown that systems integrating information on the target language side, also called string-to-tree systems, achieve the best performance on several language pairs (Bojar et al., 2014). At the same time, soft syntactic features significantly improve the translation quality of hierarchical systems (Hiero) as shown in (Marton et al., 2012; Chiang, 2010; Liu et al., 2011; Cui et al., 2010). Improving the performance of string-to-tree systems through the integration of soft syntactic constraints on the source language side is therefore an interesting task.

So far, all approaches on this topic include soft syntactic constraints into the rules of string-to-tree (Zhang et al., 2011; Huck et al., 2014) or string-to-dependency (Huang et al., 2013) systems and define heuristics to determine to what extent these constituents match the syntactic structure of the source sentence. We propose a novel way to integrate soft syntactic constraints into a string-to-tree system. We define a discriminative rule selection model for string-to-tree machine translation. We consider rule selection as a multi-class classification problem where the task is to select the correct target side of a rule given its source side as well as contextual information about the source sentence and the considered rule. So far, such models have been applied to systems without syntactic annotation on the target language side. He et al. (2008), He et al. (2010) and Cui et al. (2010) apply such rule selection models to hierarchical machine translation, Liu et al. (2008) to tree-to-string systems and Zhai et al. (2013) to systems based on predicate argument structures. When target side syntactic annotations are taken into account, the task of rule selection has to be reformulated (see Section 2) while the same type of model can be used in approaches without target annotations. This work is the first attempt to define a rule selection model for a string-to-tree system. We make our implementation publicly available as part of Moses.¹

We show in Section 2 that string-to-tree rule selection is different from the hierarchical case addressed by previous work and define our rule selection model. In Section 3 we present the training procedure before providing a proof-of-concept evaluation in Section 4.

2 Rule selection for string-to-tree SMT

2.1 String-to-tree machine translation

We present string-to-tree machine translation as implemented in Moses (which is the framework that we use). String-to-tree rules have the form $X/A \rightarrow \langle \alpha, \gamma, \sim \rangle$. On the source language side,

¹We use the string-to-tree component of Moses (Williams and Koehn, 2012; Hoang et al., 2009) in which we integrate the high-speed classifier Vowpal Wabbit http://hunch.net/~vw/.
Ces cellules présentent plusieurs caractéristiques spécifiques

Figure 1: Word-aligned sentence pairs with target-side parse.

Diverses caractéristiques importantes
Various characteristics important

Figure 2: Partial translation during decoding.

all non-terminals have the unique label $X$ while on the target language side non-terminals are annotated with syntactic labels $n_i \in N_t$. The left-hand side $X/A$ consists of source and target non-terminals. In the right hand side (rhs), $\alpha$ is a string of source terminal symbols and the non-terminal $X$. The string $\gamma$ consists of target terminals and non-terminals $n_i \in N_t$. The alignment $\sim$ is a one-to-one correspondence between source and target non-terminal symbols. String-to-tree rules are extracted from pairs of strings and trees as exemplified in Figure 1. Rules $r_1$ and $r_2$ are example rules extracted from this data.

$$(r_1) \quad X/NP \rightarrow \langle X_1 \text{ caractéristiques } X_2, JJ_1 JJ_2 \text{ caractéristiques} \rangle$$

$$(r_2) \quad X/NP \rightarrow \langle X_1 \text{ caractéristiques } X_2, NNS_1 \text{ characteristic } J_1 J_2 \rangle$$

During decoding, CYK+ chart parsing (Chapelleur et al., 1998) with cube pruning and language model scoring is performed on an input sentence such as $F$ below. Each time a rule is applied to the input sentence, candidate target trees are built. Figure 2 shows the partial translations built after the segments $\text{Diverses}$ and $\text{importantes}$ have been decoded. Given these partial translations, rule $r_1$ can be applied in a further decoding step.

$F$ (Diverses)$_{X_1}$ caractéristiques (importantes)$_{X_2}$ n’ont pas été prises en compte.
(Various)$_{X_1}$ characteristics (important)$_{X_2}$ were not considered.

2.2 String-to-tree rule selection

Rule selection is the problem of selecting the rule with the correct target side among rules with the same source side. For hierarchical machine translation (Hiero), the rule selection problem consists of choosing, among $r_3$ and $r_4$, the rule that correctly applies to $F$ ($r_3$ in our example).

$$(r_3) \quad X/X \rightarrow \langle X_1 \text{ caractéristiques } X_2, X_1 \text{ caractéristiques } X_2 \rangle$$

Rule selection models disambiguate between these rules using context information about the source sentence and the shape of the rules.

In string-to-tree machine translation, the rule selection problem is different. Because the decoding process is guided by target side syntactic annotation, partial trees built during decoding must be considered when new rules are applied. For instance, when a rule is selected to translate sentence $F$ given the partial translations in Figure 2, then the non-terminals in the target side of this rule must match the constituents selected so far. Consequently, rules $r_1$ and $r_2$ (Section 2.1) are not competing during rule selection.² Competing rules for $r_1$ would be $r_5$ and $r_6$ below.

$$(r_5) \quad X/NP \rightarrow \langle X_1 \text{ caractéristiques } X_2, JJ_1 \text{ properties } JJ_2 \rangle$$

$$(r_6) \quad X/NP \rightarrow \langle X_1 \text{ caractéristiques } X_2, JJ_1 JJ_2 \text{ features} \rangle$$

For consistency with decoding, we redefine the rule selection problem for the string-to-tree case. In this setup, it is the task of disambiguating rules with the same source side and aligned target non-terminals. As a consequence, our rule selection model (presented next) is not only normalized over the source rhs of the rules but also takes target non-terminals into account. The default rule scoring procedure for string-to-tree rules implemented in Moses uses the same normalization as we do. However, Williams and Koehn (2012) propose to normalize string-to-tree rules over the source rhs only.

²This is because their target side non-terminals are different.
2.3 Rule selection model

We denote string-to-tree rules with $X/A \rightarrow (\alpha, \gamma, \sim)$, as in Section 2.1. By $\tilde{N}t_t$, we denote target non-terminals with their alignment to source non-terminals.\footnote{For rule $r_1,r_5$ and $r_6$, $\tilde{N}t_t$ would be $JJ_1$ and $JJ_2$.} \( C(f, \alpha) \) is context information in the source sentence $f$ and the source side $\alpha$. $R(\alpha, \gamma)$ represents features on string-to-tree rules. The rule selection model estimates $P(\gamma \mid C(f, \alpha), R(\alpha, \gamma), \alpha, \tilde{N}t_t)$ and is normalized over the set $G_t$ of candidate target sides $\gamma$ for a given $\alpha$ and $\tilde{N}t_t$. The function $GTO : \alpha \rightarrow Gt_t$ generates, given the source side $\alpha$ and target non-terminals $\tilde{N}t_t$, the set $G_t$ of all corresponding target sides $\gamma$. The estimated distribution can be written as:

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

In the same fashion as (Cui et al., 2010) do for the hierarchical case, we define a global rule selection model instead of a model that is local to the source side of each rule.

To illustrate the feature templates $C(f, \alpha)$ and $R(\alpha, \gamma)$ of our rule selection model, we suppose that rule $r_1$ has been extracted from the French sentence in Figure 3. The syntactic features are:

- Does $\alpha$ match a constituent: no_match
- Type of matched constituent: None
- Lowest parent of unmatched constituent: NP
- Span width covered by $\alpha$: 3

The rule internal features are:

- Source side $\alpha$: XI_caractéristiques X2 (one feature)
- Target side $\gamma$: JJ1 JJ2_caracteristiques
- Aligned terminals in $\alpha$ and $\gamma$: caractéristiques=characteristics
- Aligned non-terminals in $\alpha$ and $\gamma$: XI=JJ1 X2=JJ2
- Best baseline translation probability: Most_Frequent

Our rule selection model is integrated in the Moses string-to-tree system as an additional feature of the log-linear model.

3 Model Training

We create training examples using the rule extraction procedure in (Williams and Koehn, 2012).\footnote{Which is based on (Galley et al., 2004; Galley et al., 2006; DeNeefe et al., 2007).} We begin by generating a rule-table using this procedure. Then, each time a rule $r : X/A \rightarrow (\alpha, \gamma, \sim)$ can be extracted from the training data, we generate a new training example. The target side $\gamma$ of the extracted rule is a positive instance and gets a loss of 0. To generate negative samples, we collect all rules $r_2, \ldots, r_n$ that have the same source language side as $r$ as well as the same aligned target non-terminals $\tilde{N}t_t$. Each of these rules is a negative example and gets a cost of 1. As an example, suppose that rule $r_1$ introduced in Section 2.1 has been extracted from the training example in Figure 1. The target side "$JJ_1 JJ_2$ characteristics" is a correct class and gets a cost of 0. The target side of all other rules having the same source side and aligned target non-terminals, such as rule $r_5$ and $r_6$, are incorrect classes.

For model training, we use the cost-sensitive one-against-all-reduction (Beygelzimer et al., 2005) of Vowpal Wabbit (VW).\footnote{Specifically, the label dependent version of Cost Sensitive One Against All which uses classification.} We avoid overfitting to training data by employing early stopping once classifier accuracy decreases on a held-out dataset.\footnote{We use the development set which is also used for MIRA tuning.}

4 Experiments

4.1 Experimental Setup

Our baseline system is a syntax-based system with linguistic annotation on the target language side (string-to-tree). We use the version implemented in the Moses open source toolkit (Hoang et al., 2009; Williams and Koehn, 2012) with standard parameters. Rule extraction is performed as in (Galley et al., 2004) with rule composition (Galley et al., 2006; DeNeefe et al., 2007). Non-lexical unary rules are removed (Chung et al., 2011) and scope-3 pruning (Hopkins and Langmead, 2010) is performed. Rule scoring is done using relative frequencies normalized over the source rhs and aligned non-terminals in the target rhs. The contrastive system is the same string-to-tree system but augmented with our rule selection model as a feature of the log-linear model.
<table>
<thead>
<tr>
<th>System</th>
<th>science</th>
<th>medical</th>
<th>news</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>34.06</td>
<td>49.87</td>
<td>18.35</td>
</tr>
<tr>
<td>Contrastive</td>
<td>34.36</td>
<td>49.57</td>
<td>18.59</td>
</tr>
</tbody>
</table>

Table 2: String-to-tree system evaluation results.

We evaluate the baseline and our global model on three domains: (1) news, (2) medical, and (3) science. The training data for news is taken from Europarl-v4. Development and test sets are from the news translation task of WMT 2009 (Callison-Burch et al., 2009). For medical we use the biomedical data from EMEA (Tiedemann, 2009). Since this is a parallel corpus only, we first removed duplicate sentences and then constructed development and test sets by randomly selecting sentence pairs. As training data for science we use the scientific abstracts data provided by Carpuat et al. (2013). Table 1 gives an overview of the corpora sizes.

Berkeley parser (Petrov et al., 2006) is used to parse the English side of each parallel corpus (for string-to-tree rule extraction) as well as for parsing the French source side (for feature extraction). We trained a 5-gram language model on the English side of each training corpus using the SRI Language Modeling Toolkit (Stolcke, 2002). We train the model in the standard way and generate word alignments using GIZA++. After training, we reduced the number of translation rules by only keeping the 30-best rules with the same source side according to the direct rule translation rule probability. Our rule selection model was trained with VW. All systems were tuned using batch MIRA (Cherry and Foster, 2012). We measured the overall translation quality with 4-gram BLEU (Papineni et al., 2002), which was computed on tokenized and lowercased data for all systems. Statistical significance is computed with the pairwise bootstrap resampling technique of Koehn (2004).

4.2 Results

Table 2 displays the BLEU scores for our experiments. On science and news, small improvements are achieved while for medical a small decrease is observed. None of these differences is statistically significant.

An analysis of the system outputs for each domain showed that the small improvements are due to the fact that in string-to-tree systems there is not enough ambiguity between competing rules during decoding. To support this conjecture, we first analyzed rule diversity by looking at the negative samples collected during training example acquisition. In a second step, we compared the results of the string-to-tree systems in Table 2 with a system where the translation rules are much more ambiguous. To this aim, we applied our approach to a hierarchical system in the same line as (Cui et al., 2010). Finally, we further tested the ability of our system to disambiguate between competing rules by training a model on the concatenation of all domains.

4.3 Analysis of Rule Diversity

The amount of competing rules during decoding can be estimated by looking at the negative samples collected for each training example. This analysis showed that the diversity of rules containing non-terminal symbols is limited. We present rules $q_1$ to $q_3$ (taken from science) to illustrate the poor diversity observed in our training examples.

\[
\begin{align*}
(q_1) & \quad X/PP \rightarrow \langle \text{à } X_1 \text{ éventail } X_3, \text{ to DT}_1 \text{ JJ}_2 \text{ variety PP}_3 \rangle \\
(q_2) & \quad X/PP \rightarrow \langle \text{à } X_1 \text{ éventail } X_3, \text{ to DT}_1 \text{ JJ}_2 \text{ range PP}_3 \rangle \\
(q_3) & \quad X/PP \rightarrow \langle \text{à } X_1 \text{ éventail } X_3, \text{ to DT}_1 \text{ JJ}_2 \text{ array PP}_3 \rangle
\end{align*}
\]

Rules $q_1$ to $q_3$ are the only rules with source side à $X_1 X_2$ évèntail $X_3$. This number is very low given that the source side contains three non-terminal symbols out of which two are adjacent. Moreover, the difference between these rules is limited to the lexical translation of évèntail. This lack of diversity is due to the constraint that competing string-to-tree rules must have the same aligned non-terminal symbols, which is taken into account when collecting negative samples. In other words, the ambiguity between translation rules in a string-to-tree system is heavily restricted by the target side syntax.

The observed lack of diversity could be minimized by allowing rules with the same source rhs to have different aligned target non-terminals. In this perspective, rule scoring should be done by normalizing over the source rhs only as in Williams and Koehn (2012). The rule selection model in Section 2.3 should then be redefined and normalized over all rules with the same source rhs. Another way to improve rule diversity would be to remove target non-terminals and use preference
Table 1: Overview of the sizes of the three domains.

<table>
<thead>
<tr>
<th>System</th>
<th>science</th>
<th>medical</th>
<th>news</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>31.22</td>
<td>48.67</td>
<td>17.28</td>
</tr>
<tr>
<td>Contrastive</td>
<td>32.27</td>
<td>49.66</td>
<td>17.38</td>
</tr>
</tbody>
</table>

Table 3: Hierarchical system evaluation results. The results in bold are statistically significant improvements over the Baseline (at confidence $p < 0.05$).

<table>
<thead>
<tr>
<th>System</th>
<th>science</th>
<th>medical</th>
<th>news</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>33.78</td>
<td>49.48</td>
<td>19.12</td>
</tr>
<tr>
<td>Contrastive</td>
<td>33.87</td>
<td>49.14</td>
<td>19.00</td>
</tr>
</tbody>
</table>

Table 4: String-to-tree system evaluation results with concatenated training data.

5 Conclusion and future work

We presented the first attempt to define a rule selection model with syntactic features for string-to-tree machine translation. We have shown that in order to be applied to the string-to-tree case, the rule selection problem must be redefined. An extensive evaluation on French-English translation tasks for different domains has shown that rule selection cannot significantly improve string-to-tree systems. An analysis of rule diversity and an empirical comparison with hierarchical rule selection indicate that the low improvements are due to the fact that the ambiguity between string-to-tree rules is too small to be improved with a rule selection model. In future work, we will use different techniques to improve the diversity of the string-to-tree rules considered during decoding in our system.

Acknowledgements

We thank all members of the DAMT team of the 2012 JHU Summer Workshop. We are especially grateful to Hal Daumé III and Ales Tamchyna for their ongoing support in the implementation of our system. We also thank Andreas Maletti for his shared expertise on tree grammars. This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 644402 (HimL) and the DFG grant Models of Morphosyntax for Statistical Machine Translation (Phase 2), which we gratefully acknowledge.

References

Alina Beygelzimer, John Langford, and Bianca Zadrozny. 2005. Weighted one-against-all. In


Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. 2007. What can syntax-based mt learn from phrase-based mt. In Proc. EMNLP.


