A Verb Lexicon Model with Source-side Syntactic Context for String-to-Tree Machine Translation

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the International Workshop on Spoken Language Translation (IWSLT)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
A Neural Verb Lexicon Model with Source-side Syntactic Context for String-to-Tree Machine Translation

Maria Nădejde¹, Alexandra Birch¹, Philipp Koehn²

¹School of Informatics, University of Edinburgh
²Department of Computer Science, Johns Hopkins University
m.nadejde@sms.ed.ac.uk, a.birch@ed.ac.uk, phi@jhu.edu

Abstract

String-to-tree MT systems translate verbs without lexical or syntactic context on the source side and with limited target-side context. The lack of context is one reason why verb translation recall is as low as 45.5%.

We propose a verb lexicon model trained with a feed-forward neural network that predicts the target verb conditioned on a wide source-side context. We show that a syntactic context extracted from the dependency parse of the source sentence improves the model’s accuracy by 1.5% over a baseline trained on a window context.

When used as an extra feature for re-ranking the n-best list produced by the string-to-tree MT system, the verb lexicon model improves verb translation recall by more than 7%.

1. Introduction

Syntax-based MT systems handle long distance reordering with synchronous translation rules. Below we show an example of a German-English synchronous rule which contains one lexical token on the source side sich which is a reflexive pronoun and several non-terminals¹:

\[ \text{root} \rightarrow \langle \text{VBZ} \text{ sich} \text{ NP} \text{ PP}, \text{ NP} \text{ VBZ} \text{ PP} \rangle \]

The non-terminals, VBZ (verb part-of-speech tag), NP (noun phrase) and PP (prepositional phrase) represent the reordering of the verb and its arguments according to the target side word order. However the rule does not contain a lexical head for the verb, the subject or the prepositional modifier. Therefore the entire predicate argument structure is translated by subsequent independent rules. The verb in particular will be translated by a lexical rule which is the equivalent of a one word phrase-pair. The language model context is also limited, and will capture at most the verb and one main argument. Due to the lack of larger source or target context the verb is often mistranslated, as shown in Figure 1. In this work we propose to improve lexical choices for verbs by learning a verb-specific lexicon model conditioned on a wide syntactic source-side context.

Several Discriminative Word Lexicon (DWL) models with source-side features have addressed the problem of word sense disambiguation in phrase-based MT [1, 2, 3]. However this is the first work that addresses specifically the problem of verb translation in string-to-tree systems. We train a verb-specific lexicon model since verbs have the most outgoing dependency relations, are central to semantic structures and therefore would benefit most from a source-side syntactic context.

The proposed verb lexicon model is trained with a feed-forward neural network (FFNN) which, unlike DWL models, allows parameter sharing across target words and avoids exploding feature spaces. Previous lexicon models trained with FFNN [4] using global source-side context were inefficient to train and did not scale to large vocabularies. We avoid scaling problems by choosing the context which is most relevant for verb prediction in a pre-processing step, from the source-side dependency structure.

Our results show that the verb lexicon model with global syntactic context outperforms the baseline model with local window context by 1.5%. Furthermore when used as a feature for reranking, the verb lexicon model improves verb translation precision by up to 2.7% and recall by up to 7.4% at the cost of a small (less than 0.5%) decrease in BLEU score.

2. Related work

Several approaches have been proposed to improve word sense disambiguation (WSD) for machine translation by integrating a wider source context than is available in typical translation units.

For phrase-based MT one such approach is to learn a discriminative lexicon model as a maximum-entropy classifier which predicts the target word or phrase conditioned on a highly dimensional set of sparse source-side features. [5] train a classifier for each source phrase and use features engineered for Chinese WSD to choose among available phrase translations. [6] propose a similar model that uses target-side features and that shares parameters across all source

---

¹String-to-tree translation rules have generic (X) non-terminal labels on the source-side that correspond one-to-one with syntactic non-terminal labels on the target side.
phrases. [1] introduced the Discriminative Word Lexicon (DWL) which models target word selection independently of which phrases are used by the MT model. The DWL is a binary classifier that predicts whether a target word should be included or not in the translation, conditioned on the set of source words. [2] extend the DWL with target-side context and bag-of-n-gram features aimed at capturing the structure of the source sentence. [3] extend the work of [2] with other source-side structural features such as dependency relations.

For syntax-based MT, discriminative models have been used to improve rule selection [7, 8, 9]. The rule selection involves choosing the correct target side of a synchronous rule given a source side and other features such as the shape of the rule and the syntactic structure of the source span covered by the rule. [8] proposes a global discriminative rule selection model for hierarchical MT which allows feature sharing across all rules and which incorporates a wider source context such as words surrounding the source span. However, the model only disambiguates between rules with the same source side. Considering that hierarchical rule tables are much larger than phrase tables, the discriminative rule selection models are much more expensive than the discriminative lexicon models.

The aforementioned DWL models train a separate classifier for each target word or phrase. The classifier parameters are not shared across target words and the feature combinations are not learned but generated through cross-products of feature templates. Joint translation models trained with feed forward neural networks (FFNN) [10] address these problems however they are efficiently trained only on local context. [4] proposes a joint model with global context similar to the DWL but trained with FFNN. However the resulting network is very large and inefficient to train and therefore the model does not scale to large vocabularies.

Our work is similar to [3] as we select relevant source context following the dependency relations between the verb and its arguments. However we take advantage of parameter sharing and avoid the problem of exploding feature space by training our model with a FFNN. Different from [4] we are able to incorporate more global context by taking advantage of the syntactic structure of the source sentence. We train a verb specific lexicon model with the knowledge that verbs have the most outgoing dependency relations, are central to semantic structures and therefore would benefit most from a source-side syntactic context. We train a lexicon model and not a rule selection model as we are trying to address the problem of lexical translation of verbs in string-to-tree systems. Moreover by predicting only the target verb we can simplify the prediction task and train a smaller model.

3. Verb Translation Analysis

In this section we determine the extent to which verb translation is a problem for syntax-based MT systems. We estimate the impact of a verb lexicon model through the percentage of verbs that would benefit from source-side context and the increase in verb translation recall that can be gained from n-best lists. We present an analysis of verb translation in syntax-based models for the German to English language pair. This language pair is challenging for machine transla-
A second problem for verb translation is that synchronous rules may translate the verb independently of its arguments. Table 1 shows that 20.8% of the verbs are translated without context with lexical rules which are the equivalent in phrase-based terms of one word phrase-pairs. When translating verbs with lexical rules the system relies only on language model context to disambiguate the verb. However the language model context might become available only in later stages of bottom-up chart-based decoding, when larger synchronous rules are applied to connect and reorder the verb and its arguments. To address this problem we propose a verb lexicon model that uses a wide source-side context to predict the target verb.

An interesting class of German verbs are those with separable particles which are moved at the end of the sentences for present tense or imperative. For example the verbs *ausgehen* (to go out) and *fortgehen* (to leave) have the root *gehen* (to walk). However the particles *aus* and *fort* separate from the root and change its meaning, which leads to a specific type of translation errors.

We continue to evaluate the tree-to-string system in terms of verb translation recall. The translation recall shown in Table 2 is computed over the 11,161 instances of main source verbs aligned to target verbs.

<table>
<thead>
<tr>
<th>source</th>
<th>token</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-best</td>
<td>45.54</td>
<td>53.14</td>
</tr>
<tr>
<td>1000-best</td>
<td>72.87</td>
<td>79.24</td>
</tr>
<tr>
<td>rule table</td>
<td>91.85</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Verb translation recall for 1-best translation, 1000-best lists and rule table computed over verbs from newstest2013-2015.

Verb translation recall is only 45.5% at token level for the 1-best output of the syntax-based system. However verb recall in the 1000-best list is much higher, at 72.87%. This result indicates that better translation options are available and re-scoring these options could result in improved 1-best verb translation recall. Furthermore by looking at the target side of all the verb translations in the rule table we can see that the reference translation is available in almost 92% of the cases.

Finally we compare the reference translations and the system translations in terms of their rank among all translation options. For this purpose we order the translation options for each of the source verbs according to the direct translation probability \( p(\text{target} | \text{source}) \). For each source verb we compute the rank of the corresponding reference translation and that of the translation produced by the syntax-based system. We can see in Table 3 that the reference translations have rank 1 only 50.71% of the time compared to 65.48% for the system translations. Since the correct translation of the verb is less probable than the selected one we are dealing with modeling errors. Re-scoring only the top
In recent years, a number of scientists have studied the links [between ... and cancer].

Table 2: Example of window and syntactic context extracted for the source verb *untersucht* (studied). With both word and lemma factors for the syntactic context, the network’s input size is 16.

<table>
<thead>
<tr>
<th>Reference:</th>
<th>Window context</th>
<th>Syntactic context</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the last years, a number of scientists have studied the links [between ... and cancer]</td>
<td>und Krebs untersucht</td>
<td>source verb parent dependents pp modifier subcat particle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>untersucht haben Wissenschaftler Zusammenhang &lt;null&gt; in Jahren pp_subj_obja &lt;null&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>untersuchen haben Wissenschaftler Zusammenhang &lt;null&gt; in Jahr - -</td>
</tr>
</tbody>
</table>

10 translation options could improve the translation model accuracy from 50.7% to 68.25%.

<table>
<thead>
<tr>
<th>source</th>
<th>rank = 1</th>
<th>rank &lt; 5</th>
<th>rank &lt; 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference</td>
<td>50.71</td>
<td>56.30</td>
<td>68.25</td>
</tr>
<tr>
<td>system</td>
<td>65.48</td>
<td>73.90</td>
<td>84.87</td>
</tr>
</tbody>
</table>

Table 3: Percentage of reference and system translations of verbs in newstest2013-2015, that are ranked in the rule table below a threshold.

4. Verb Lexicon Model

In the previous section we have shown that string-to-tree MT systems translate verbs with low recall and accuracy. Better translation options can be found in the 1000-best lists. However, at least 20% of verbs are scored without contextual information.

In this section we propose a verb lexicon model that uses source side context to predict the target verb. Both the source word sequences and the source syntactic structure are readily available at early stages of decoding. In contrast, target side context for verbs such as their arguments becomes available at later stages of decoding, when larger synchronous rules are applied. Moreover the target syntactic structure generated during decoding is not sufficiently accurate for extracting arguments of the target verb. While similar lexicon models have been proposed in the literature [1, 2, 3], this work explores whether a source syntactic context is more informative for predicting target verbs than a window context. We propose a verb specific model since verbs have more arguments and longer syntactic dependencies than other words and therefore would benefit from a wider source-side context. Our verb lexicon model is a feed-forward neural network trained with the NPLM toolkit [14].

We first show that verbs are harder to predict cross-lingually than other words for the German-English language pair. For this purpose we train a generic lexicon model that takes as input a 5-word window centered on the source word of interest and outputs the corresponding target word. The generic model is trained on all words from WMT15 parallel data and evaluated on either all words from newstest2013-2015 test sets or on the subset of 11,161 main verbs selected as described in the previous section. Table 4 shows that the generic lexicon model performs worse at predicting target verbs: perplexity is higher, 26.20 for verbs compared to 23.62 for all words, and accuracy is lower, 43.67 for verbs compared to 50.62 for all words. This reinforces our argument that we need a verb specific lexicon model.

<table>
<thead>
<tr>
<th></th>
<th>perplexity</th>
<th>acc@1</th>
<th>acc@5</th>
<th>acc@15</th>
</tr>
</thead>
<tbody>
<tr>
<td>all words</td>
<td>23.62</td>
<td>50.62</td>
<td>70.51</td>
<td>78.47</td>
</tr>
<tr>
<td>verbs only</td>
<td>26.20</td>
<td>43.67</td>
<td>67.88</td>
<td>78.69</td>
</tr>
</tbody>
</table>

Table 4: Perplexity and accuracy of the generic lexicon model reported over all words and over verbs only, on newstest2013-2015.

4.1. Syntactic Context

In order to improve accuracy of predicting target verbs as well as to train the models more efficiently, we learn a specialized verb lexicon model. The network receives a fixed number of input tokens extracted from the source sentence...
Next we explore whether a source-side syntactic context is more informative for predicting the target verb than a window context. Since the syntactic context is extracted from the source sentence we can include most of the verb’s dependents, in particular the core arguments that carry most semantic information relevant to verb disambiguation. Diathesis alternations, represented with subcategorization features, have been used to induce verb classes in a monolingual setting [15, 16]. Therefore we also provide the verb lexicon model with a feature encoding the subcategorization frame.

From the dependency parse of the source sentences we extract the following syntactic context: the parent of the verb, one prepositional modifier, up to three other dependents and the verb particle, if any. We create a subcategorization token by concatenating the dependency relations of all verb dependents. In order to reduce sparsity of the data we add the lemma of each word in the syntactic context. If all types of syntactic context are considered the network will receive 16 input tokens. We show an example of source syntactic context for a verb in Figure 2. In this example there are 9 pieces of context, out of which 7 have both a word and lemma factor, resulting in a total of 16 inputs for the neural network.

4.2. Experimental Setup and Evaluation

The models are trained with the NPLM toolkit [14] implementing a feed-forward neural network. We used 200 dimensions both for the input embeddings and for the single hidden layer. Both the input and output vocabularies consist of the 500,000 most frequent types. The input vocabulary is shared for words and lemmas. When adding the subcategorization information we increase the input vocabulary by 80,000. We use the "rectifier" activation function, a batch size of 256, and train for at most 25 iterations.

We train the models on all the parallel training data available at WMT15 and a development set of 2,000 sentences for early stopping of training. The models are evaluated in terms of perplexity and accuracy over the verbs extracted from newstest2013, newstest2014, newstest2015\(^3\). The data is described in Table 6. The source side of the parallel data is parsed with dependency relations using ParZU [12] and the target side is tagged with part-of-speech labels using TreeTagger [13].

Table 5 shows the performance of different models. The accuracy of the verb lexicon model trained with a 5-word window context is 50.57\%, compared to 43.67\% the accuracy of the generic lexicon model reported in Table 4. This result shows that training a verb-specific model is beneficial. In Table 3 we showed that the direct translation probability predicts the correct translation for 50.71\% of the verbs that have a translation in the rule table. The prediction of the verb lexicon model with window context matches the reference translation in 50.57\% of the cases, however its top 5 accuracy is 76.27\% compared to only 56.30\% for the direct translation probability.

Increasing the window context size to 7 words does not improve performance of the verb lexicon model. In contrast providing a syntactic context of similar size as input to the network results in a lower perplexity and higher accuracy. Adding the lemma factor helps for both types of context in terms of perplexity, however the accuracy is higher only for the syntactic context. Surprisingly the subcategorization information did not help. The reason might be that the target-side of some synchronous rules, such as the example in Section 1, already encode the subcategorization information for the target verb. Finally, adding the particle as separate input increases the accuracy leading to a total improvement of 1.5\% over the baseline window context.

In the next section we investigate whether the verb lexicon model is able to improve translation quality by integrating the model as an additional feature for re-reranking machine translation output.

---

Table 6: Number of sentences in the training, tuning and test sets.

<table>
<thead>
<tr>
<th>context</th>
<th>factors</th>
<th>size</th>
<th>perplexity</th>
<th>acc@1</th>
<th>acc@5</th>
<th>acc@15</th>
</tr>
</thead>
<tbody>
<tr>
<td>window</td>
<td>word</td>
<td>5</td>
<td>27.81</td>
<td>50.57</td>
<td>76.27</td>
<td>85.04</td>
</tr>
<tr>
<td>window</td>
<td>word</td>
<td>7</td>
<td>27.98</td>
<td>50.57</td>
<td>75.55</td>
<td>85.03</td>
</tr>
<tr>
<td>window</td>
<td>word, lemma</td>
<td>10</td>
<td>27.20</td>
<td>50.54</td>
<td>75.90</td>
<td>85.42</td>
</tr>
<tr>
<td>syntactic</td>
<td>word</td>
<td>7</td>
<td>26.49</td>
<td>51.21</td>
<td>76.26</td>
<td>85.36</td>
</tr>
<tr>
<td>syntactic</td>
<td>word, lemma</td>
<td>14</td>
<td>24.99</td>
<td>51.46</td>
<td>77.12</td>
<td>85.83</td>
</tr>
<tr>
<td>syntactic</td>
<td>word, lemma, subcat</td>
<td>15</td>
<td>25.16</td>
<td>51.54</td>
<td>76.83</td>
<td>85.82</td>
</tr>
<tr>
<td>syntactic</td>
<td>word, lemma, subcat, particles</td>
<td>16</td>
<td>24.84</td>
<td>51.99</td>
<td>77.54</td>
<td>85.96</td>
</tr>
</tbody>
</table>

\(^3\)In Section 3 we compared reference and system translations of 11,161 source verbs. The remaining verbs up to 11,211 were discarded since there was no corresponding system translation in the 1-best output.
BLEU and METEOR scores are reported over newstest2015 (2,169 sentences and 3,002 reference verbs) with standard deviation shown from 3 runs of MERT.

<table>
<thead>
<tr>
<th>context</th>
<th>factors</th>
<th>BLEU (dev)</th>
<th>BLEU (test)</th>
<th>METEOR (dev)</th>
<th>METEOR (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-</td>
<td>26.18 ± 0.0</td>
<td>26.10 ± 0.0</td>
<td>29.95 ± 0.0</td>
<td></td>
</tr>
<tr>
<td>+ window</td>
<td>word</td>
<td>-0.13 ± 0.08</td>
<td>-0.39 ± 0.26</td>
<td>-0.13 ± 0.14</td>
<td></td>
</tr>
<tr>
<td>+ dependency</td>
<td>word, lemma, subcat</td>
<td>-0.06 ± 0.05</td>
<td>-0.22 ± 0.12</td>
<td>-0.07 ± 0.08</td>
<td></td>
</tr>
<tr>
<td>+ dependency</td>
<td>word, lemma, subcat, particles</td>
<td>-0.13 ± 0.06</td>
<td>-0.37 ± 0.19</td>
<td>-0.14 ± 0.06</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Results of re-ranking 1000-best lists with different configurations of the verb lexicon model as an additional feature. Precision, recall and F1 scores for verb translation are reported over newstest2015 (2,169 sentences and 3,002 reference verbs) with standard deviation shown from 3 runs of MERT.

<table>
<thead>
<tr>
<th>context</th>
<th>factors</th>
<th>Precision (token)</th>
<th>Precision (lemma)</th>
<th>Recall (token)</th>
<th>Recall (lemma)</th>
<th>F1 (token)</th>
<th>F1 (lemma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-</td>
<td>56.91 ± 0.0</td>
<td>65.18 ± 0.0</td>
<td>47.86 ± 0.0</td>
<td>54.83 ± 0.0</td>
<td>51.99 ± 0.0</td>
<td>59.56 ± 0.0</td>
</tr>
<tr>
<td>+ window</td>
<td>word</td>
<td>+1.95 ± 0.66</td>
<td>+2.04 ± 0.31</td>
<td>+7.45 ± 0.42</td>
<td>+8.54 ± 0.72</td>
<td>+5.04 ± 0.32</td>
<td>+5.57 ± 0.28</td>
</tr>
<tr>
<td>+ dependency</td>
<td>word, lemma, subcat</td>
<td>+2.44 ± 0.80</td>
<td>+2.39 ± 0.68</td>
<td>+7.14 ± 0.76</td>
<td>+7.5 ± 1.08</td>
<td>+5.00 ± 0.09</td>
<td>+5.42 ± 0.28</td>
</tr>
<tr>
<td>+ dependency</td>
<td>word, lemma, subcat, particles</td>
<td>+2.70 ± 0.85</td>
<td>+2.5 ± 0.72</td>
<td>+7.93 ± 0.40</td>
<td>+7.76 ± 0.06</td>
<td>+5.34 ± 0.56</td>
<td>+5.53 ± 0.32</td>
</tr>
</tbody>
</table>

5. Machine Translation Evaluation

Our baseline system for translating German into English is the Moses string-to-tree toolkit implementing GHKM rule extraction [17, 18, 19]. The string-to-tree translation model is based on a synchronous context-free grammar (SCFG) that is extracted from word-aligned parallel data with target-side syntactic annotation. The system was trained on all available data provided at WMT15 [26] on 1000-best lists. We report evaluation scores over the newstest2015 data set (2169 sentences, 3002 verbs).

We integrate the verb lexicon model in reranking by adding two new features scores in addition to the baseline features:

- A counter for the source verbs translated by the n-best hypothesis.
- Verb lexicon model score aggregated over all main verbs.

The weights for the new feature scores and for the baseline features are re-tuned using MERT on the tuning set. We run MERT three times and for each set of weights we ranked the machine translation output.

Table 7 shows average BLEU [27] and METEOR [28] scores, as well as the standard deviation for the three different tuning runs. When adding the verb lexicon model there is a small decrease in both scores: less than 0.4% for BLEU and less than 0.2% for METEOR. Table 8 shows average precision, recall and F1 scores for verb translation, as well as the standard deviation for the three different tuning runs. On average the verb lexicon model improves precision up to 2.7%, recall up to 7.4% and F1 scores up to 5.3% at token level. The models with syntactic context perform slightly better in terms of precision compared to the models with window context, but not in terms of recall. This result motivates future work on analyzing how verb recall is affected by tuning feature weights towards BLEU, a precision based metric. We consider a 7% gain in verb translation recall to be more important than the small decrease in BLEU and METEOR scores since verbs are key pieces in semantic structures. Perhaps an even stronger verb lexicon model is needed in order to out-weight choices that only improve fluency. Model coverage could be improved by making predictions for predicative nouns and model accuracy could be improved by conditioning on target context. Based on our analysis in Section 3, choosing from the n-best list allows for significant verb recall improvements, however this improvement may come at a cost to BLEU.

In Figure 1 we give examples of correct verb translations produced by re-ranking the 1000-best list with the verb lexi-
Und so geht das Leben, anders als das vieler anderer, für uns weiter.

So life goes on for us, unlike for so many.

And so is the life, unlike many others, for us.

The verb lexicon model is able to produce the correct translation particle as a preposition to a noun-phrase. The verb lexicon model is able to produce a better translation accused, however this affects the choice of other translation rules. As a result the second verb angegriffen (attacked) and its prepositional modifiers are incorrectly reordered in the translation.

6. Conclusions

We proposed a verb lexicon model to address the problem of low verb translation recall of string-to-tree MT systems. The model is trained with a feed-forward neural network that predicts the target verb conditioned on a wide source-side context. We have shown that a syntactic context extracted from the dependency structure of the source sentence improves model accuracy by 1.5% over the baseline window context.

The verb lexicon model was used as an extra feature for re-ranking the output of a baseline string-to-tree MT system. The model improved verb translation precision by up to 2.7% and recall by up to 7.4% at the cost of a small (less than 0.5%) decrease in BLEU score. Surprisingly the verb lexicon model trained on syntactic context improved only verb translation precision and not recall, as compared to the less accurate model trained on window context. This result motivates future work on analyzing how verb recall is affected by tuning feature weights towards BLEU, a precision based metric.
7. Acknowledgements
We thank the anonymous reviewers for their comments and suggestions. This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreements 644402 (HimL) and 645452 (QT21).

8. References


