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

Maria Nădejde[1], *Alexandra Birch*[1], *Philipp Koehn*[2]

[1]School of Informatics, University of Edinburgh
[2]Department 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-terminals[1]:

$$root \rightarrow \langle VBZ\ sich\ \ NP\ PP,\\ NP\ \ VBZ\ \ PP\ \rangle$$

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

[1] String-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.

a)	Source	Die Kongress Abgeordneten haben einen Gesetzesvorschlag **eingebracht** ,								
		um die Organisation von Gewerkschaften als Bürgerrecht zu etablieren .								
	Reference	Congressmen have **proposed** legislation to protect union organizing as a civil right .								
	Baseline	Congressmen have **tabled** a bill to establish the organization of trade unions as a civil right .								
	Verb Lexicon	Congressmen have **introduced** a bill to establish the organization of trade unions as a civil right .								
	Syntactic context	source verb	parent	dependents			pp modifier		subcat	particle
	word:	**eingebracht**	haben	Kongress	Gesetzesvorschlag	etablieren	<null>	<null>	subj_obja_neb	<null>
	Translation rule	$VP \rightarrow \langle haben\ NP\ eingebracht\ um\ S\ ,\ have\ tabled\ NP\ S \rangle$								
b)	Source	die Ankläger **legten** am Freitag dem Büro des Staatsanwaltes von Mallorca Beweise								
		für Erpressungen durch Polizisten und Angestellte der Stadt Calvia **vor** .								
	Reference	the claimants **presented** proof of extortion by policemen and Calvia Town Hall civil servants								
		at Mallorca's public prosecutor's office on Friday .								
	Baseline	the prosecutor **went** to the office of the prosecutor of Mallorca Calvi evidence of extortion								
		by police officers and employees of the city on Friday .								
	Verb Lexicon	the prosecutor **presented** evidence of extortion by police officers and employees of the city								
		on Friday the office of the prosecutor of Mallorca Calvi before .								
	Syntactic context	source verb	parent	dependents			pp modifier		subcat	particle
	word:	**legten**	<null>	Ankläger	Büro	Staatsanwaltes	am	Freitag	subj_pp_objd_obja_pp_pp_avz	vor
	Translation rule	$VP \rightarrow \langle legten\ \hat{VP}\ ,\ went\ \hat{VP} \rangle$								
		$PP \rightarrow \langle NP\ vor\ ,\ to\ NP \rangle$								

Figure 1: Examples of correct verb translation produced by re-ranking the 1000-best list with the verb lexicon model.

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 translation because German allows the word order of Subject-Verb-Object to be both SVO and OVS, while in English it is always SVO. German also allows verbs to appear in different positions: in perfect tense the main verb appears at the end of the sentence and some verbs have separable particles that are placed at the end of the sentence. Syntax-based models handle such long distance reordering with synchronous rules which may translate verbs independently of their arguments.

The string-to-tree system used for this analysis is trained on all available data from WMT15 [11] and is described in more detail in Section 5. The evaluation test set consists of newstest2013, newstest2014 and newstest2015 totaling 8,172 sentences. 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].

Firstly, we present in Table 1 a breakdown of counts at token level for verbs identified in the source sentences. Verbs were first identified by their part-of-speech label and then the dependency relations were used to distinguish between auxiliary verbs (except modals) and main verbs. Main verbs represent 73.2% of all verbs while only 20.0% are auxiliary verbs. The other 6.8% of words labeled as verbs are either modals or can not be identified as either auxiliaries or main verbs.

	count	percentage
source verbs	23,492	100.0
|__ auxiliary verbs	4,689	20.0
|__ misaligned verbs	934	3.9
|__ main verbs	17,210	73.2
|__ particle verbs	1,589	6.7
|__ **target verbs**	**11,161**	**47.5**
|__ misaligned verbs	2,850	12.1
|__ modals + other	1,593	6.8
|__ lexical rules	4,905	20.8

Table 1: Breakdown of source verb categories in newstest2013-2015. Token level counts.

The first problem for verb translation is misalignment, verbs aligned with at least one comma or not aligned at all, which breaks the constraints for synchronous rule extraction. A total of 16% of verbs are misaligned, with 20% of auxiliaries[2] and 16.5% of main verbs being misaligned. In this work we will focus on translation of main verbs as they carry the semantic information. In order to avoid the problem of misalignment, we restrict the training and evaluation data to source verbs that align with target verbs, as identified by their part-of-speech label. This leaves us with a total of 11,161 verbs for which we can evaluate the impact of a verb lexicon model.

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.

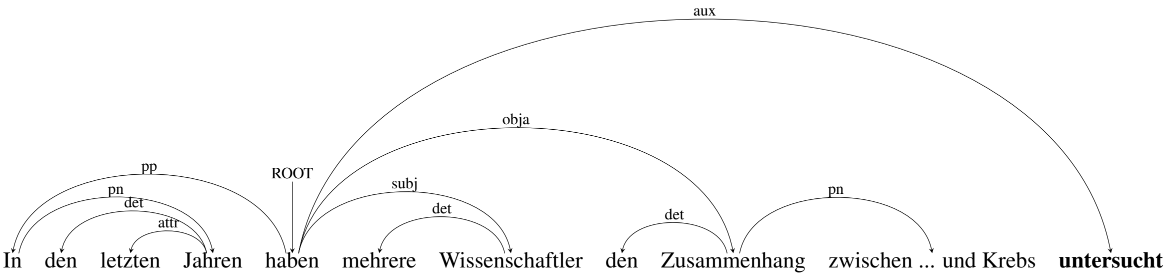
source	token	lemma
1-best	45.54	53.14
1000-best	72.87	79.24
rule table	91.85	-

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 $p(target|source)$. 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

[2] Not all German auxiliaries need to be translated into English. For example a different form of past tense can be used. *habe gegessen* translates as *ate*.


aux
obja
pp
pn
det
attr
ROOT
subj
det
det
pn
In den letzten Jahren haben mehrere Wissenschaftler den Zusammenhang zwischen ... und Krebs **untersucht**

Reference:	In recent years , a number of scientists have **studied** the links [between ... and cancer]								
Window context	und	Krebs	**untersucht**	$</s>$	$</s>$				
Syntactic context	source verb	parent	dependents			pp modifier		subcat	particle
word:	**untersucht**	haben	Wissenschaftler	Zusammenhang	<null>	in	Jahren	pp_subj_obja	<null>
lemma:	untersuchen	haben	Wissenschaftler	Zusammenhang	<null>	in	Jahr	-	-

Figure 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.

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

source	rank = 1	rank < 5	rank < 10
reference	50.71	56.30	68.25
system	65.48	73.90	84.87

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.

	perplexity	acc@1	acc@5	acc@15
all words	23.62	50.62	70.51	78.47
verbs only	26.20	43.67	67.88	78.69

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

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

Table 5: Evaluation of different configurations of the verb lexicon model. The size column indicates the number of inputs to the neural network. Token level verb prediction accuracy is reported over newstest2013-2015.

and predicts a target verb.

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.

	Train	Tune	Test
sentences	4,472,694	2,000	8,172
verb tokens	5,945,637	2,419	11,211

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

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.

[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		METEOR
context	factors	dev	test	test
Baseline	-	$26.18_{\pm 0.0}$	$26.10_{\pm 0.0}$	$29.95_{\pm 0.0}$
+ window	word	$-0.13_{\pm 0.08}$	$-0.39_{\pm 0.26}$	$-0.13_{\pm 0.14}$
+ dependency	word, lemma, subcat	$-0.06_{\pm 0.05}$	$-0.22_{\pm 0.12}$	$-0.07_{\pm 0.08}$
+ dependency	word, lemma, subcat, particles	$-0.13_{\pm 0.06}$	$-0.37_{\pm 0.19}$	$-0.14_{\pm 0.06}$

Table 7: Results of re-ranking 1000-best lists with different configurations of the verb lexicon model as an additional feature. 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.

		Precision		Recall		F1	
context	factors	token	lemma	token	lemma	token	lemma
Baseline	-	$56.91_{\pm 0.0}$	$65.18_{\pm 0.0}$	$47.86_{\pm 0.0}$	$54.83_{\pm 0.0}$	$51.99_{\pm 0.0}$	$59.56_{\pm 0.0}$
+ window	word	$+1.95_{\pm 0.66}$	$+2.04_{\pm 0.31}$	$+7.45_{\pm 0.42}$	$+8.34_{\pm 0.72}$	$+5.04_{\pm 0.32}$	$+5.57_{\pm 0.28}$
+ dependency	word, lemma, subcat	$+2.44_{\pm 0.80}$	$+2.39_{\pm 0.68}$	$+7.14_{\pm 0.76}$	$+7.8_{\pm 1.08}$	$+5.09_{\pm 0.09}$	$+5.42_{\pm 0.28}$
+ dependency	word, lemma, subcat, particles	$+2.70_{\pm 0.89}$	$+2.5_{\pm 0.72}$	$+7.36_{\pm 0.40}$	$+7.76_{\pm 0.06}$	$+5.34_{\pm 0.56}$	$+5.53_{\pm 0.32}$

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.

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 [4] [11]. The number of sentences in the training and tuning sets are shown in Table 6. The English side of the parallel corpus is parsed using the Berkeley parser [20, 21].

We use the rule extraction parameters proposed by [22] for German-English: *Rule Depth = 5, Node Count = 20, Rule Size = 5*. At decoding time we give a high penalty to glue rules and allow non-terminals to span a maximum of 50 words. We train a 5-gram language model on all available monolingual data [5] using the SRILM toolkit [23] with modified Kneser-Ney smoothing [24] for training and KenLM [25] for language model scoring during decoding. The feature weights were tuned using the Moses implementation of MERT[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 reranked 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-

[4] http://www.statmt.org/wmt15/translation-task.html

[5] target side of the parallel corpus, the monolingual English News Crawl, Gigaword and news-commentary

	Source	Und so **geht** das Leben , anders als das vieler anderer , für uns **weiter** .
	Reference	So life **goes on** for us unlike for so many .
a)	Baseline	And , unlike many others , life **goes on** for us .
	Verb Lexicon	And so **is** the life , unlike many others , for us .
	Syntactic context	source verb parent dependents pp modifier subcat particle
	word:	**geht** <null> und so Leben <null> <null> S_koord_adv_subj <null>
	Translation rule	$VBZ \rightarrow \langle geht\ ,\ goes \rangle$
		$\hat{V}P \rightarrow \langle PP\ weiter\ ,\ on\ PP \rangle$
	Source	Webster wird darber hinaus **vorgeworfen** , am 4. Mai 2014 eine zweite Frau im Golf View Hotel in Naim im schottischen Hochland angegriffen zu haben .
b)	Reference	Webster is then **charged** with attacking a second woman at the Golf View Hotel in Nairn in the Highlands on May 4 , 2014 .
	Baseline	Webster is also **alleged** to have attacked a second woman in Naim's Golf View Hotel in the Scottish Highlands on 4 May 2014 .
	Verb Lexicon	Webster is also **accused** of being a second wife in the Golf View Hotel on 4 May 2014 in Naim attacked in the Scottish Highlands .
	Syntactic context	source verb parent dependents pp modifier subcat particle
	word:	**vorgeworfen** wird Webster haben <null> darüber hinaus S_objd_pp_subjc <null>
	Translation rule	$VBN \rightarrow \langle vorgeworfen\ ,\ alleged \rangle$
		$VP \rightarrow \langle VBN\ ,\ VP\ zu\ haben\ ,\ VBN\ to\ have\ VP \rangle$

Figure 3: Examples of translations produced by re-ranking the 1000-best list with the verb lexicon model that are worse than the 1-best translations.

con model.

In example a), the verb *eingebracht (proposed)* is translated incorrectly by the baseline system as *tabled*, which implies rejecting a bill. On the last row of the example we show the synchronous translation rule used by the baseline system to translate the verb *eingebracht*. The rule correctly re-orders the noun-phrase *a bill* and the verb, as English objects should come after the verb. However the lexical choice for the verb is made without knowledge of the lexical head of the object. The re-ranked translation *introduced* is correct, and the verb lexicon model prefers this translation because the words *Kongress* and *etablieren* appear in the syntactic context.

In example b), the verb *vorlegten (presented)* is translated incorrectly by the baseline system as *went*. This happens because the verb has a separable particle *vor* which is moved at the end of the sentence. The string-to-tree system is not able to find a rule that would make such a long distance reordering. Instead it translates the verb with two rules that are disconnected. The first rule translates the verb without any other context. The second rule incorrectly attaches the verb particle as a preposition to a noun-phrase. The verb lexicon model is able to produce the correct translation *presented* as the particle *vor* appears in the syntactic context.

In Figure 3 we give examples where the translations produced by re-ranking the 1000-best list with the verb lexicon model are worse than the 1-best translations.

In example a) the verb *geht weiter* is correctly translated by the baseline system as *goes on* but incorrectly translated by the verb lexicon model as *is*. The parser is not able to identify *weiter* as dependent of the source verb, therefore the verb lexicon model has limited context and gives a lower score to *goes* and a higher score to *is*. The wrong choice for the verb causes the resulting translation to have worse word order.

In example b) the verb *vorgeworfen* is incorrectly translated by the baseline system as *alleged*. 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

[1] A. Mauser, S. Hasan, and H. Ney, “Extending statistical machine translation with discriminative and trigger-based lexicon models,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, ser. EMNLP ’09, 2009, pp. 210–218.

[2] J. Niehues and A. Waibel, “An mt error-driven discriminative word lexicon using sentence structure features,” in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 2013, pp. 512 – 520.

[3] T. Herrmann, J. Niehues, and A. Waibel, “Source discriminative word lexicon for translation disambiguation,” in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, 2015.

[4] T. Ha, J. Niehues, and A. Waibel, “Lexical translation model using a deep neural network architecture,” in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT 2014)*, 2014.

[5] M. Carpuat and D. Wu, “Improving statistical machine translation using word sense disambiguation,” in *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 61–72.

[6] A. Tamchyna, A. M. Fraser, O. Bojar, and M. Junczys-Dowmunt, “Target-side context for discriminative models in statistical machine translation,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1704 – 1714.

[7] F. Braune, N. Seemann, and A. Fraser, “Rule selection with soft syntactic features for string-to-tree statistical machine translation,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1095–1101. [Online]. Available: https://aclweb.org/anthology/D/D15/D15-1129

[8] F. Braune, A. Fraser, H. Daumé III, and A. Tamchyna, “A framework for discriminative rule selection in hierarchical moses,” in *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 92–101. [Online]. Available: http://www.aclweb.org/anthology/W16-2210

[9] Q. Liu, Z. He, Y. Liu, and S. Lin, “Maximum entropy based rule selection model for syntax-based statistical machine translation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 89–97. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613715.1613729

[10] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, “Fast and Robust Neural Network Joint Models for Statistical Machine Translation,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, MD, USA, June 2014, pp. 1370–1380.

[11] O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi, “Findings of the 2015 workshop on statistical machine translation,” in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September 2015, pp. 1–46.

[12] R. Sennrich, M. Volk, and G. Schneider, “Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis,” in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2013*, Hissar, Bulgaria, 2013, pp. 601–609.

[13] H. Schmid, “Probabilistic part-of-speech tagging using decision trees,” in *International Conference on New Methods in Language Processing*, Manchester, UK, 1994, pp. 44–49.

[14] A. Vaswani, Y. Zhao, V. Fossum, and D. Chiang, “Decoding with Large-Scale Neural Language Models Improves Translation,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, USA, 2013, pp. 1387–1392.

[15] L. Sun and A. Korhonen, “Improving verb clustering with automatically acquired selectional preferences,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, ser. EMNLP ’09, 2009, pp. 638–647.

[16] S. S. im Walde, “Experiments on the automatic induction of german semantic verb classes,” *Comput. Linguist.*, vol. 32, no. 2, pp. 159–194, June 2006.

[17] M. Galley, M. Hopkins, K. Knight, and D. Marcu, “What’s in a translation rule?” in *Proceedings of Human Language Technologies: Conference of the North*

American Chapter of the Association of Computational Linguistics, ser. HLT-NAACL '04, 2004.

[18] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer, "Scalable inference and training of context-rich syntactic translation models," in *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Morristown, NJ, USA, 2006, pp. 961–968.

[19] P. Williams and P. Koehn, "Ghkm rule extraction and scope-3 parsing in moses," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 2012, pp. 388–394.

[20] S. Petrov, L. Barrett, R. Thibaux, and D. Klein, "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*, ser. ACL-44, 2006, pp. 433–440.

[21] S. Petrov and D. Klein, "Improved Inference for Unlexicalized Parsing," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, Rochester, New York, April 2007, pp. 404–411.

[22] M. Nadejde, P. Williams, and P. Koehn, "Edinburgh's Syntax-Based Machine Translation Systems," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August 2013, pp. 170–176.

[23] A. Stolcke, "SRILM – an Extensible Language Modeling Toolkit," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, vol. 3, Sept. 2002.

[24] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Harvard University, Tech. Rep., 1998.

[25] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, ser. WMT '11, Edinburgh, Scotland, 2011, pp. 187–197.

[26] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03, Morristown, NJ, USA, 2003, pp. 160–167.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[28] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, ser. StatMT '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 228–231. [Online]. Available: http://dl.acm.org/citation.cfm?id=1626355.1626389